DOCUMENT RESUME

ED 095 212 TM 003 885

AUTHOR	Lown, Donald E.
TITLE	Standardized Tests: A Handbook for Administration and Use.
INSTITUTION	Greece Central School District, Rochester, N.Y.
PUB DATE	[74]
NOTE	60p.
EDRS PRICE	MF-$0.75 HC-$3.15 PLUS POSTAGE
DESCRIPTORS	Data Analysis; *Guidelines; School Districts; *Standardized Tests; Testing; *Testing Programs; *Test Interpretation; *Test Results
IDENTIFIERS	*Greece Central School District; Rochester

ABSTRACT

The purpose of this handbook is to provide information and guidelines for using data from standardized tests. The handbook consists of six sections: (1) explains the rationale for district-wide testing; (2) discusses a district testing program; (3) presents guidelines for testing which describes procedures for administering standardized tests; (4) interpretation and uses of standardized test data is presented by discussions about the purpose for standardized testing, norms, test scores, types of report formats, uses of individual student data, and uses of group data; (5) summary; and (6) glossary of terms. (MLP)

ED 095212

STANDARDIZED TESTS

A Handbook For Administration And Use

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

TM 003 885

GREECE CENTRAL SCHOOL DISTRICT
Research and Evaluation
Donald E. Lown, Supervisor
Jo Ellen Sherman, Research Assistant

GREECE CENTRAL SCHOOL DISTRICT
1790 Latta Road
Rochester, N.Y.
P.O. Box 7197
North Greece, N.Y. 14515

Text and Drawings by
Donald E. Lown, Supervisor of Research and Evaluation
with the editorial and technical assistance of
Jo Ellen Sherman, Research Assistant

TABLE OF CONTENTS

PROLOGUE 5
DISTRICT TESTING RATIONALE 9
- PERSPECTIVE 11
- EVALUATION 11
- AN OBJECTIVE 12
- SOME IMPORTANT POINTS 12
- CONCLUSION 13

THE DISTRICT TESTING PROGRAM 15
- WHY TEST? 17
- WHAT ARE STANDARDIZED TESTS? 17
- BACKGROUND ABOUT DEVELOPING OUR PROGRAM 17
- THE TESTS WE USE 19

GUIDELINES FOR TESTING 21
- BEFORE TESTING 23
- DURING (OR AT THE TIME OF) TESTING ... 24
- AFTER TESTING 33

INTERPRETATION AND USES OF DATA . 35
- PURPOSE 37
- NORMS 37
- TEST SCORES 38
  - Raw scores 38
  - Percentages 38
  - Standard scores 38
  - Percentiles 39
  - Stanines 39
  - Grade equivalents 40
- TYPES OF REPORT FORMATS 41
  - Class lists 41
  - Frequency distributions and statistical data ... 43
  - Item analysis 45
  - Less commonly used forms 47
- USES OF INDIVIDUAL STUDENT DATA 48
  - What to look for 48
    - *Look for the usual* 48
    - *Look for the unusual* 49
    - *Don't emphasize small differences* 49
    - *Don't expect to discover something new about every student* 49
  - Using students' scores 49
    - *Instructional applications* 49
    - *Bases for selection* 49
    - *Guidance and counseling* 50
- USES OF GROUP DATA 51
  - Group statistics 51
    - *The median, quartiles and percentiles* 51
  - Item analysis 52
  - Group achievement across subtests 56
  - Correlation with other data 56

SUMMARY 57

GLOSSARY OF TERMS 63

NOTE: Bold face entries appearing throughout the text identify terms which appear in the glossary.

PROLOGUE

TESTING IS A TECHNIQUE FOR OBTAINING INFORMATION

PROLOGUE

The purpose of this handbook is to provide information and guidelines for using data from Standardized Tests. Many facets to this topic exist. I urge you to become familiar, not only with the contents of this handbook, but also with topics, details, and points of view regarding standardized tests that are not to be found here.

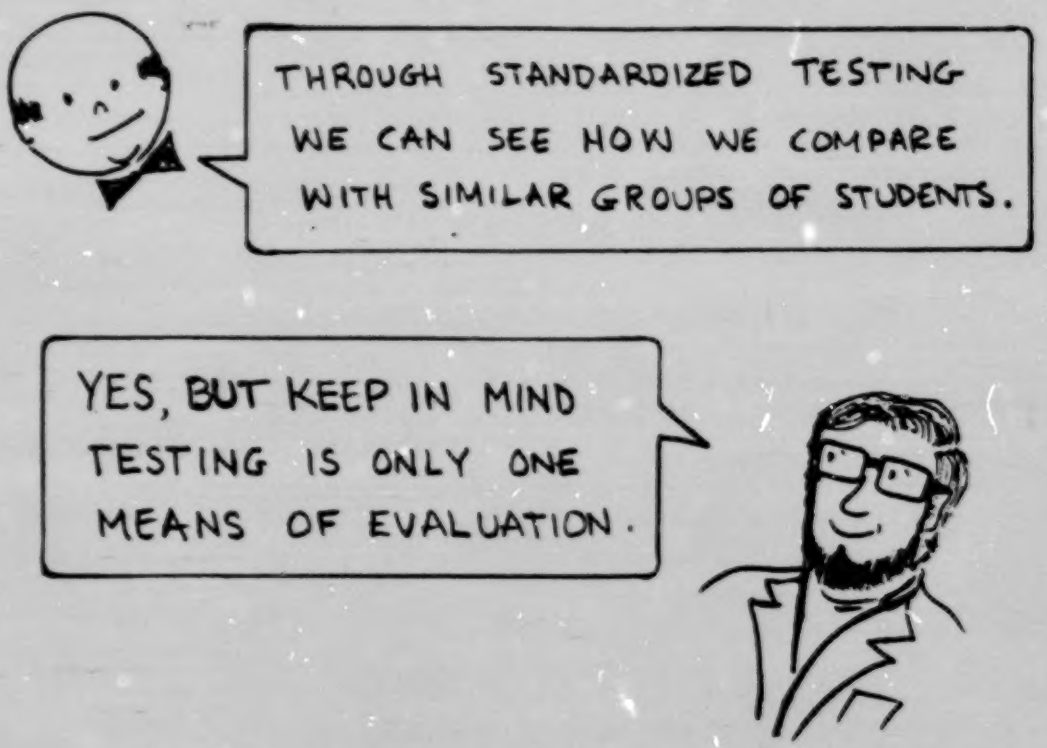
Testing is neither a policy, a set of principles or beliefs, nor a program. Testing is a technique for obtaining information. As a tool, testing is neutral; it serves the ends of the user. The better the user understands the strengths and limitations of test data, the greater are the opportunities to maximize benefits from testing. The less well one understands this tool, the greater are the risks of inadvertent misuse of test data.

One of the great premises upon which our society is constitutionally based allows that all men are created equal. It is one of the unfortunate misapplications of our time that there has been a thoughtless transfer of the concept of equality of individuals under the law to learning theory. People are inherently different, unique, with needs that must be addressed in relationship to, and with value placed upon, those differences. Recognizing that schools have historically focused and probably will continue to primarily focus upon a relatively narrow segment of the human dimension, how then should we view human inequalities as they relate to learning? The first step, I believe, is to hold human diversity in high regard.

Initially, the use of standardized tests seems contrary to the nurture of human differences. The very concept of standardization looks to the common, average, or normative behavior. Such reference, however, is counter to fostering individual differences only if normative descriptions, intentionally or unintentionally, become goals rather than tools.

Donald E. Lown
Supervisor of Research and Evaluation

DISTRICT TESTING RATIONALE


THROUGH STANDARDIZED TESTING WE CAN SEE HOW WE COMPARE WITH SIMILAR GROUPS OF STUDENTS.
YES, BUT KEEP IN MIND TESTING IS ONLY ONE MEANS OF EVALUATION.

Standardized testing is only one means of learning about the fulfillment of our goals. That this handbook speaks to this one means of evaluation is not to overlook nor minimize the importance of other means. Subjective judgments of experienced and/or knowledgeable individuals may often yield better evaluation than the so called objective information. "Objective" data from standardized testing are the focus of this handbook. It is our intention that such data be seen as valuable, kept in perspective, and used appropriately to the benefit of students.

DISTRICT TESTING RATIONALE

Discussions in this section expose our beliefs, assumptions and rationale that relate to standardized testing.

PERSPECTIVE

The purpose of schools in American Society is the topic of much debate, with a lack of agreement. As an institution, the schools are variously charged with the transmission of our culture; being agents of change for the future, teaching "basic skills" while leaving the formation of values to the home and/or church; helping students to build their value systems, and so on! In such a milieu of diversity, serving many masters with ambiguity of goal expectations will remain reality. How does this translate into daily work?

A large and diverse school district is essentially a microcosm of the society. District goals and philosophy are an amalgam from all of us who comprise the district: parents, teachers, students, administrators, taxpayers and support staff with many tasks. Regardless of what may be written, for each of us individually and collectively as a district, the philosophic base upon which our policies are based is a matter of day to day dynamics. In general, once written, policies tend to become static and thus obsolete as conditions change. This should neither disturb us nor prevent us stating policies and courting change.

EVALUATION

Evaluation is an integral part of all activity. The term evaluation is used here in its broadest sense. It is unfortunate that evaluation often has a negative connotation, for such limited meaning prevents recognition of the pervasive part that evaluation plays in our daily behaviors. Decisions are not made without evaluation having occurred.

Evaluation, as a process, causes effects on the situations being evaluated. As a means of evaluation, then, standardized testing, although it may be considered a neutral tool, has its effect; whether the effect is good or bad depends upon the purposes to which information from testing is applied. An evaluation process, of which testing is only one aspect, may provide useful information and yet interfere in some way with the realization of broad educational goals. None the less when results from testing are an integral part of attaining specified goals (as opposed to an outside influence) negative effects are minimized and offset by the usefulness of the information.

While not every effect of standardized testing is positive, no thoughtful consideration of a district's educational program is complete without data from such testing. Our goal is to reap as many benefits from an integrated testing program as possible while keeping to a minimum any adverse effects.

It is assumed that there are many important educational objectives that are not measured by standardized tests. Test related data are only one means of evaluation. That this discussion focuses on test data merely delimits what is here, it in no way implies greater importance or value relative to other evaluative bases.

AN OBJECTIVE

A District objective is to achieve and maintain a level of student accomplishment consistent with District, State and/or National comparative data.

Seeking to meet this objective does not supplant previously established learning objectives. Further, all efforts and outcomes in this pursuit must be compatible with the goal of "humanizing" the educational environment and the relationship between teacher and student. Keeping the foregoing considerations in mind, the public we serve are none the less entitled to know how well their schools are carrying out the education of their children. One means at our disposal for evaluating and evolving the quality of education is to compare our results with those of a larger population through the use of standardized tests, Regents examinations and other related exams (ACT, SAT, etc.).

SOME IMPORTANT POINTS

All behaviors should manifest a regard for the dignity and value of each individual. Building a healthy self-concept for each student is at least as important as building proficiency in basic skills. Evaluative feedback to the individual in a positive, humane way helps to prevent uncertainty and misunderstanding about oneself. **Achievement** information should be used as a basis for future learning activities, and not in a threatening, hurtful, judgmental evaluation of an individual's worth.

As part of the regard for each individual, the District must provide the opportunity, assistance and environment for students to achieve successfully, each to his own level.

Data, relevant and valid, are important to decision making. Student achievement scores on **standardized tests** constitute one means of assessing how the District is achieving a part of its educational responsibility. Such scores are important for the student's academic record, the community's perception of their schools, and information for administration of school programs.

Standardized tests differ in pupose from Assessment tests. Standardized achievement tests are given primarily to determine the progress of student groups for purposes of program planning and the monitoring of achievement at the school and District levels.

Standardized tests are those that are given in accordance with some very specific administration requirements of time, statement of directions, etc. Adhering to such standards is an attempt to control all pertinent variables that could affect student achievement aside from the student's own capability. Thus one can compare the achievement of any group of students to the achievement of the normative population. Standardized tests are designed to test a fairly wide range of skills in order to relate students' overall achievement.

Assessment type tests, on the other hand, are **criterion referenced.** That is, specifically stated objectives are tested to determine whether or not a student has mastered a particular skill. Instructional planning is a direct and immediate result. Assessment tests, as differentiated from standardized tests, are one aspect of the diagnostic process used for planning programs of study for individual students; these are based on local objectives. Where local objectives are not established, standardized test data at an individual level are necessary for instructional decisions; as

assessment testing increases fulfilling the information need at the instructional level, the need for standardized test information decreases in frequency. Standardized testing serves as a program monitoring connection to the "outside world."

Standardized achievement test data should affect, but should not determine the curriculum. As an example, assume that learning to be proficient in the use of the metric system is part of our curriculum. If test results show that students are weak in their knowledge of the metric system (compared to specified objectives or to normative data), it would be appropriate for this information to have a significant effect upon what is being done relative to study of the metric system. Since harness making is not a component of our curriculum, scoring very poorly on questions from a standardized test would not warrant bringing harness making into the curriculum. Standardized test information, like the tail of a dog, should indicate quite a bit about its owner, but the tail should not wag the (curricular) dog.

All students at each grade level tested should take the same test each year. Group descriptive statistics that are to be compared with some reference data depend for their **validity** and comparability upon the data from both groups being derived from the same test (or an equated alternate form of the test). To follow District ends requires that 6th graders this year take the same test as did last year's 6th grade class. Exclusion of a portion of a grade level population would distort group statistics making relationships uncertain.

Normative figures are not goals, but reference points from which to measure one's position. Norms for a standardized test are descriptions of the average (normal) achievement of students in the standardization sample for that specific test. As such they may be used as guides for comparing results of other groups. Note that the average achievement of a standardization group may not be satisfactory for local expectations. This may be due to differences in characteristics of the local population compared to the standardization group, or simply to different levels of aspiration--different goals.

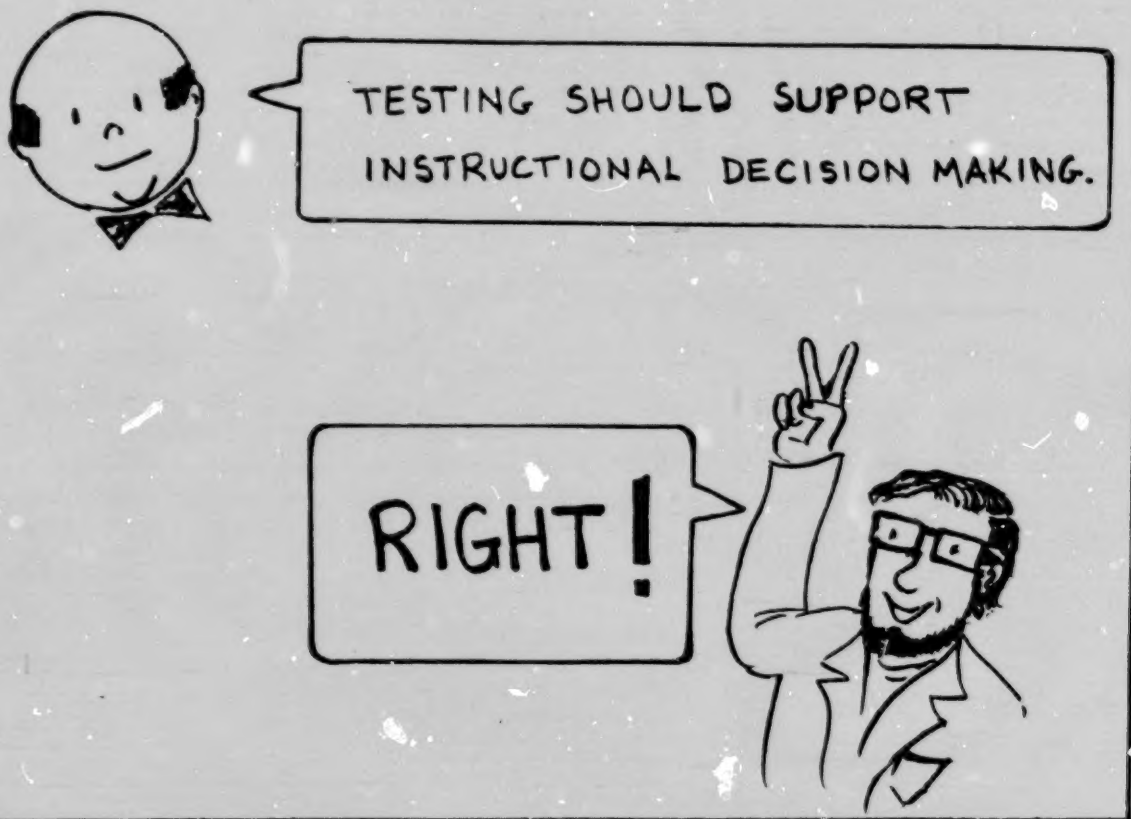
Inasmuch as our District has a fairly large student population (approximately 1000 per grade level), group statistics would be expected to remain quite stable from year to year. Based on the heterogenity of our population and the lack of any large group that would distort from a normally distributed population, there is no reason to expect that achievement by our students as a District would be much different from standardization normative values.

CONCLUSION

The foregoing statements relate to our current program. As needs change, the standardized testing program must change in accommodation. Recognizing both the dynamic nature of curriculum and the need for stability imposed by the desire for trend/comparative statistics, the current standardized testing program was adopted for a 4 year period. During the 4th year (1974-75) a study will be made to see whether changes in the program (the tests used, frequency of testing, population tested, etc.) may be necessary.

■■■

THE DISTRICT TESTING PROGRAM



THE DISTRICT TESTING PROGRAM

WHY TEST?

Students take tests and educators administer tests in order to learn something about the status of students' knowledge, skill competencies, sttitudes or whatever. Assuming that any or every test provides a valid measure, probably the major factor affecting the usefulness of test data lies in having a clearly defined purpose for such data. A single test may be perfectly appropriate for one purpose but grossly unfair and inappropriate for some other purpose. Purpose is extremely important.

Two types of testing are carried out on a districtwide level: Assessment testing measures how well a student achieves on specific local curricular objectives. Examples are reading and math assessment processes (RAP and MAP). Standardized testing provides an anchor to the rest of the world by allowing comparison of our students' achievement to that of a large normative population on a broad, general basis.

WHAT ARE STANDARDIZED TESTS?

The term "standardized" relates to those test instruments that have:

1) very particular time limits
2) specific instructions to be adhered to
3) been administered to a large population for the purpose of establishing **"norms"**
4) (ideally) undergone rigorous development to insure statistical **validity** (the good tests on the market have been so developed!).

"Standardization" derives from having standards. Unfortunately the concept of standards is too often misapplied; the standards become confused, even equated, with the norms (viewed as goals) rather than being rules (standard ways) for administering the test. When tests are standardized on large populations, any other group or individual taking the test - - following the same standard procedures - - can compare results with those of the large population (the norms). Note well that norms are descriptions of the **achievement** of typical students in a large population, not goals nor standards that set expectations (for further discussion see page 37 under Interpretation and Uses of Data).

BACKGROUND ABOUT DEVELOPING OUR PROGRAM

District needs and available testing materials were extensively studied by a committee of some 35 people over a three year period. The committee involved teachers, guidance counselors, principals, administrators and parents. The considerable work of this group led to submission of recommendations to the administration accompanied by these statements:

> **"It is accepted that we cannot expect 100% agreement districtwide with the structure of any testing program due to differing perspectives in a controversial area. None the**

less, the committee submits the attached recommendation after careful and extensive study in confidence and with consensus that these constitute a meaningful program that represents our needs for assisting the learner.

Formulation of the Standardized Testing Program is based on three premises:

1. The Standardized Testing Program, as part of an overall evaluation program, exists to provide information to decision-makers who exist at various levels, as examples:

   a. students: knowledge of strengths and weaknesses helps a student (and his parents) to better decide on his own purposes for learning and planning future studies.

   b. teachers: knowledge of individual student status and progress assists the teacher to better decide on and guide subsequent steps in a student's learning program.

   c. principals: knowledge of the progress of groups of students supports decisions on allocation of resources, scheduling of facilities, needs of teaching staff and needs of students.

   d. district administrators: knowledge of student achievement supports decisions in the areas of program evaluation, curriculum development and overall program management.

2. The Standardized Testing program should consist of only those tests (in type and frequency) that fulfill clearly defined needs. Thus, misuse of data and over-testing will be minimized.

3. Standardized Testing should complement other means of evaluation by being:

   a. curriculum embedded. That is, the tests used and thus the data received should be part of our instructional program. As such, test results serve a **formative** function; they assist in forming serial steps in the learning/instructional process.

   b. tools for program evaluation. In this urage, standardized tests serve a **summative** function. Data are used to sum up on an annual basis, the achievement of individuals and groups as measured by the test.

Some rationale associated with these premises will help to clarify the attached recommendations.

There seems to be no justification for collecting data simply because we ought to do something. Those data that are collected should clearly support instructional decision making. This leads to a minimal schedule of regularly administered standardized tests that is essentially independent of school organization. While the timing and location of testing within the K-12 spectrum might need minor adjustments as organization patterns change, the systematic measurement of students and programs should carry on regardless of specific school organization. For example, a change to a 7-9 junior high school or a 6-8 middle school organization ought not disrupt the continuity of the Standardized Testing Program.

In view of many limitations inherent in normative test data, standardized testing is not appropriate for showing short term learning gains. Thus, summative evaluations via standardized tests should not be more frequent than

annually and the greater emphasis should be reserved for longer durations such as from 3rd to 6th to 9th grades for trend overviews. Formative applications require more frequent data collection.

Formative testing occurs when the results of testing are used to help formulate the next instructional step. Formative data must derive from a test that is highly correlated with the curriculum. Criterion oriented tests, such as those based on specific objectives as in math and reading assessment, as well as teacher made tests, are the most frequently used formative instruments. Standardized tests serve to relate the content of local curricula to a national, state, or regional base by allowing comparison of local results to those of the larger normative population. On an annual basis, standardized test results may serve a formative function in program evolution.

An evaluation program consists of the planned intertwining of criterion (formative) and normative (standardized) testing."

Acceptance of the Standardized Testing Committee's recommendations led to adoption of the program as outlined in Figures 3.1 and 3.2.

THE TESTS WE USE

The choice of the Metropolitan Achievement Test (MAT) was based on several considerations.

a) The MAT is current (1970 edition); both content and format have been recently revised.
b) A study of the MAT subsections in Reading, Language and Math showed these to be highly related to our curriculum.
c) The MAT utilizes a continuous **standard score** scale -- which allows administration of "out-of-normal level" tests to individuals and theoretically still retains the capability of relating results to the reference group. This feature is important to us in cases where regular testing indicates that individual retesting would be advisable. More accurate, yet relatable indices may thus be reached.
d) The MAT is obtained from a company with a reliable reputation for both materials and services.

The addition of the Iowa Test of Basic Skills (ITBS) work-study skills section to our program is to add a dimension not covered by the MAT. The ITBS work-study skills test measures skills in the areas of map reading, reading graphs and tables, and using reference materials.

The Otis-Lennon Mental Ability Test provides a measure of students' competencies that are known to correlate well with academic achievement. This allows comparison of results on at least two different measures--the ability, or **aptitude**, test and the achievement test.

The New York State Pupil Evaluation Program (PEP) tests in reading and math are mandated statewide for all 3rd, 6th, and 9th graders.

Optional tests are provided when need is identified by a teacher, administrator, or subject supervisor. Materials are obtained from the Research and Evaluation Office. Examples of optional tests include the Primer level MAT for late Kindergarten or first grade, Spache Diagnostic Reading Test, Nelson-Denny Silent Reading Test, MAT science subsection, Ohio Vocational Interest Survey, etc. Many hundreds of test materials are on the market. In circumstances of clearly identified need, any item is theoretically available. Requests for the use of extra-program test materials should be made through the appropriate subject supervisor.

Standardized Testing Program

The tests that comprise the testing program are:

1) METROPOLITAN ACHIEVEMENT TEST (MAT) - sections on Word Knowledge, Word Analysis, Reading, Language, Spelling, Math Computation, Math Concepts, and Math Problem Solving.

2) IOWA TEST OF BASIC SKILLS (ITBS) - work-study skills section only.

3) OTIS-LENNON MENTAL ABILITY TEST. (OLMAT)

4) NEW YORK STATE PUPIL EVALUATION PROGRAM (PEP) - tests in Reading and Mathematics.

5) OTHERS-OPTIONAL - several other tests are used on an individual basis for specific instructional and/or evaluative needs.

Figure 3.1

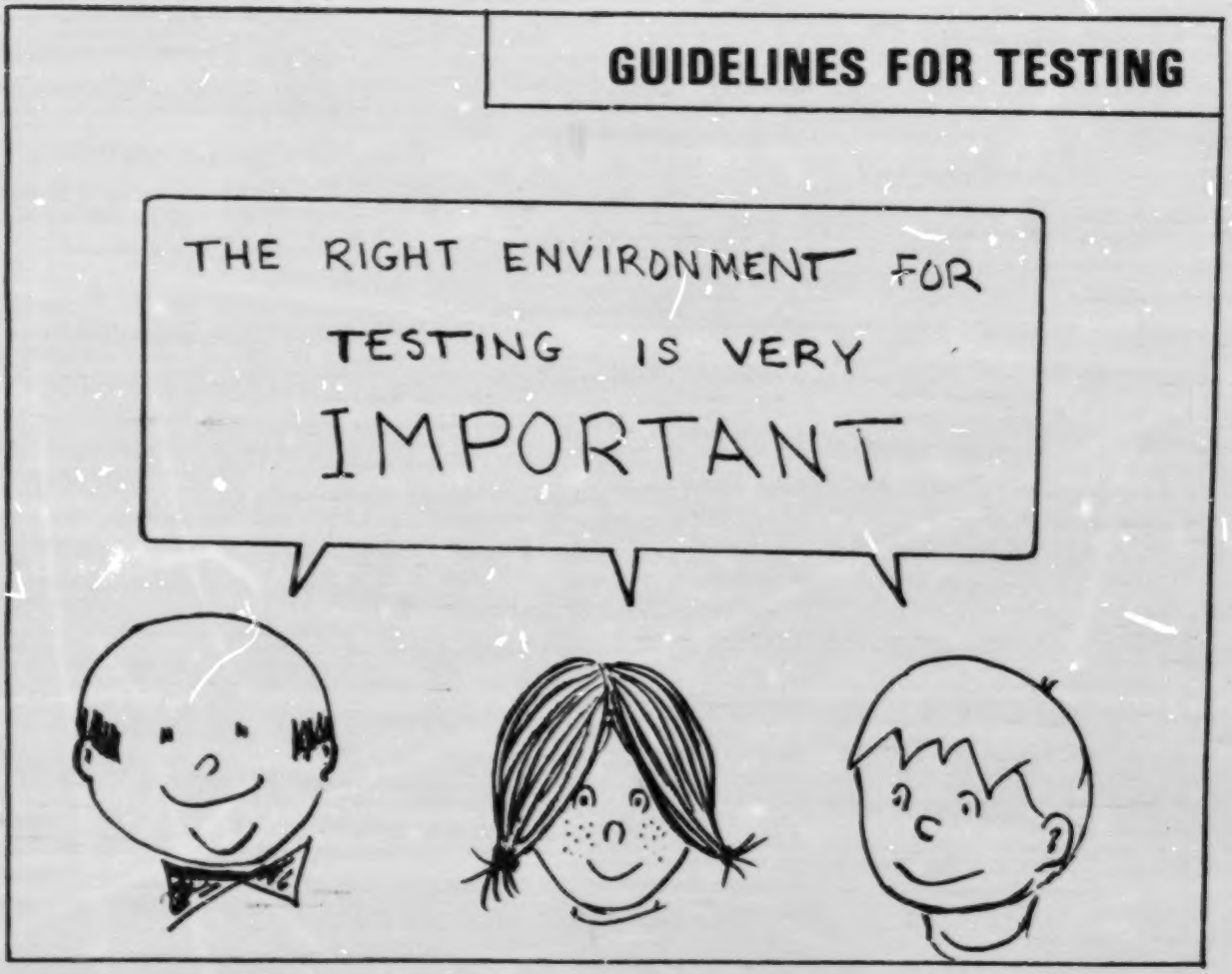
Standardized Testing Schedule

The testing schedule is as follows:

Grade	1	MAT - Primer Battery - optional usage	
		MAT - Primary I Battery	Spring
Grade	2	MAT - Primary II Battery	Spring
Grade	3	NYS-PEP - Reading and Math Tests	Fall
		OLMAT	Fall
		MAT - Elementary Battery	Spring
		ITBS - Work-Study Skills Section only	Spring
Grade	4	MAT - Elementary Battery	Spring
Grade	5	MAT - Intermediate Battery	Spring
Grade	6	NYS-PEP - Reading and Math Tests	Fall
		OLMAT	Fall
		MAT - Intermediate Battery	Spring
		ITBS - Work-Study Skills Section only	Spring
Grade	7	MAT - Advanced Battery	Spring
Grade	8	MAT - Advanced Battery	Spring
Grade	9	NYS-PEP - Reading and Math Tests	Fall
		MAT - Advanced Battery	Spring
		ITBS - Work-Study Skills Section only	Spring

Figure 3.2

GUIDELINES FOR TESTING



GUIDELINES FOR TESTING

Inasmuch as there are various situations under which testing may be initiated, any single set of "rules" must be fairly general. If one is considering administration of districtwide testing, obviously it is incumbent upon someone with districtwide coordination responsibility to assure that all materials are available, that schedule and mechanics are known, etc. If one is considering testing a single class, the teacher may bear the burden of initiating materials procurement (through the Research and Evaluation Office). If the testing of a single individual is being considered, the person(s) having responsibility for preparation may be different from the other situations, but the task of being prepared remains.

Establishing an appropriate environment for testing is a pervasive responsibility; it is held at all levels within the school system. The person taking a test has the right to the physical, psychological and emotional setting that optimizes the opportunity to do his valid best.

Assuming the foregoing, several specific responsibilities fall to one who would administer a **standardized test.** In order to obtain useful comparative data from standardized testing, it is imperative that appropriate procedures for administration be followed. These procedures are presented here in three categories which include what to do (1) before testing, (2) during testing, and (3) after testing.

BEFORE TESTING

The teacher (person administering the test) is responsible for action around each of the following points:

1. Know the purpose for testing. Whether testing is externally or self-initiated, the person who administers the test, as well as the person(s) who will use the data, should understand the reason for the testing. The person taking the test should also understand the purpose for testing. (Section II discusses rationale behind our districtwide testing program.)
2. Know the test.
   a. Familiarize yourself with the content and makeup of the test.
   b. Read the publisher's directions for test administration in order to be completely acquainted with procedures, time restrictions, etc. for each part of the test.
   c. Know the test and answer sheet in order to respond quickly and surely to questions raised by the person taking the test.
3. Prepare students for testing - set up a good test climate.

   Much can be done to assist those to be tested. Well in advance of the test date, the student should know
   a. that he will be tested,

b. why he will be tested,
c. how it will benefit him to be tested,
d. what will be done with test results.

Students should be comfortable psychologically as well as physically. A certain readiness tension, tone, or set, is good so that the student addresses the test situation with seriousness of purpose. This should not, however, be a frustration nor a fearful tension.

Since the purpose behind different standardized test vary, the mind-set that a student should have in being prepared for a test also varies from one situation to another. It is important to distinguish here two types of tests or test purposes. (1) Standardized tests used within the district: such tests are tools to learn more, both for the individual student's benefit, and for student benefit collectively through program analysis. Students should know that, to serve as a tool, testing offers a benefit rather than posing a threat. (2) Tests with rigid administration requirements such as scholarship qualifying exams (PSAT, RSE, ACT, SAT): toward such tests students should know that individual results carry important reward values. Helping the student to see such tests in the perspective of his own goals should put purpose in place of threat.

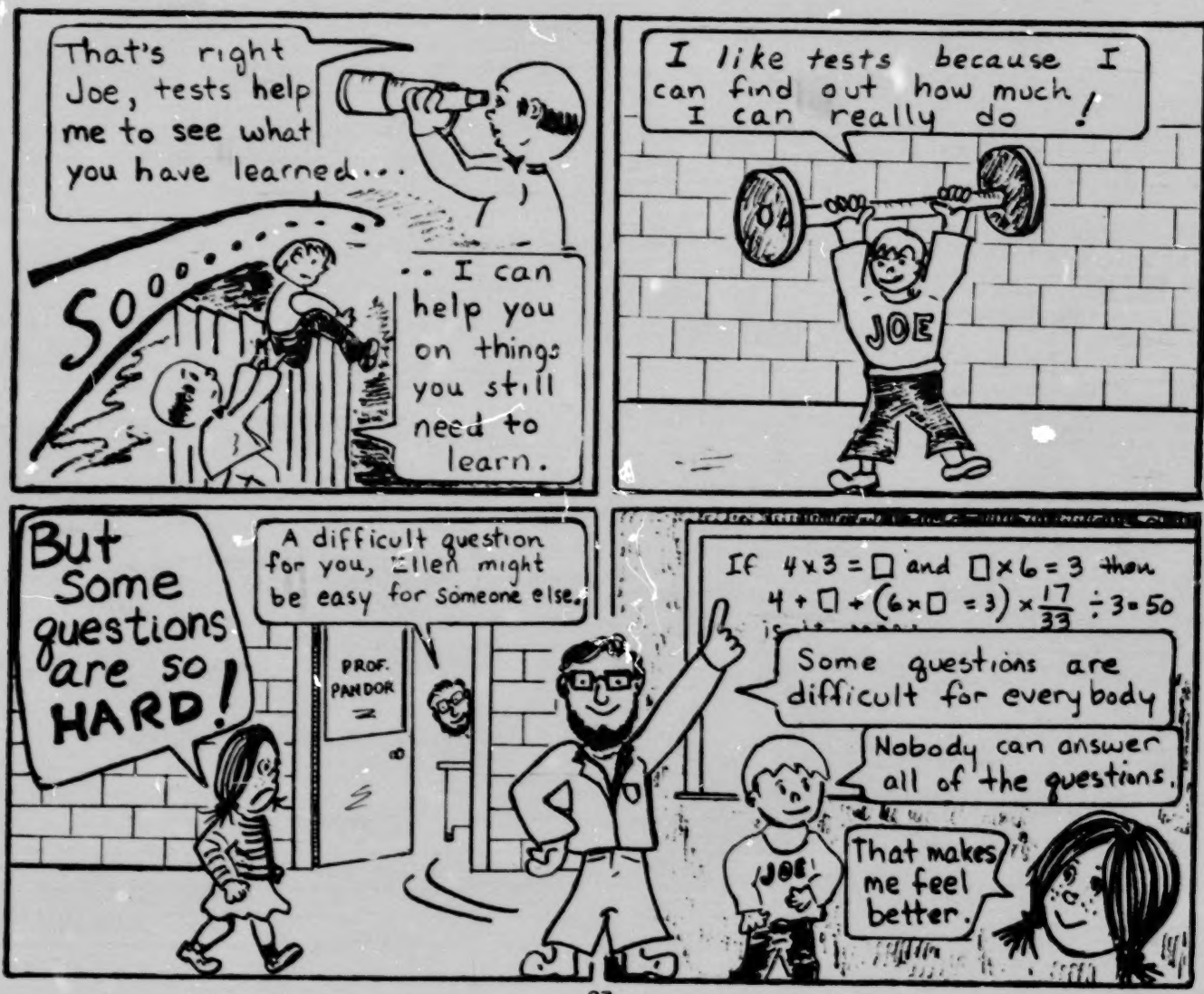
4. Select the appropriate test.
   Test selection for any student depends upon the purpose for testing. In a districtwide testing, in order to be able to obtain data about any particular grade level, all students at a level should take the same test. For students whose day to day functioning deviates considerably from the suggested test for a grade level, a different test might be administered on an individual follow-up basis to general testing. Such a testing would serve the teacher's purpose of getting a more appropriate reading of the student's individual achievement.
5. Prepare the testing schedule.
   Have the exact days and times for the test(s) to be administered determined well in advance of the testing date. Make sure that administrators, other teachers, and students are acquainted with the testing schedule.
6. Prepare testing materials.
   Have a sufficient quantity of all testing materials (tests, answer sheets, pencils, scratch paper if needed, demonstration materials, and whatever else may be necessary) on hand and well organized. Organization of all materials should be completed at least a day in advance of the testing so that no stress or confusion exists due to management problems as testing begins. A relaxed, self-assured demeanor on the part of the test administrator carries over to the student.

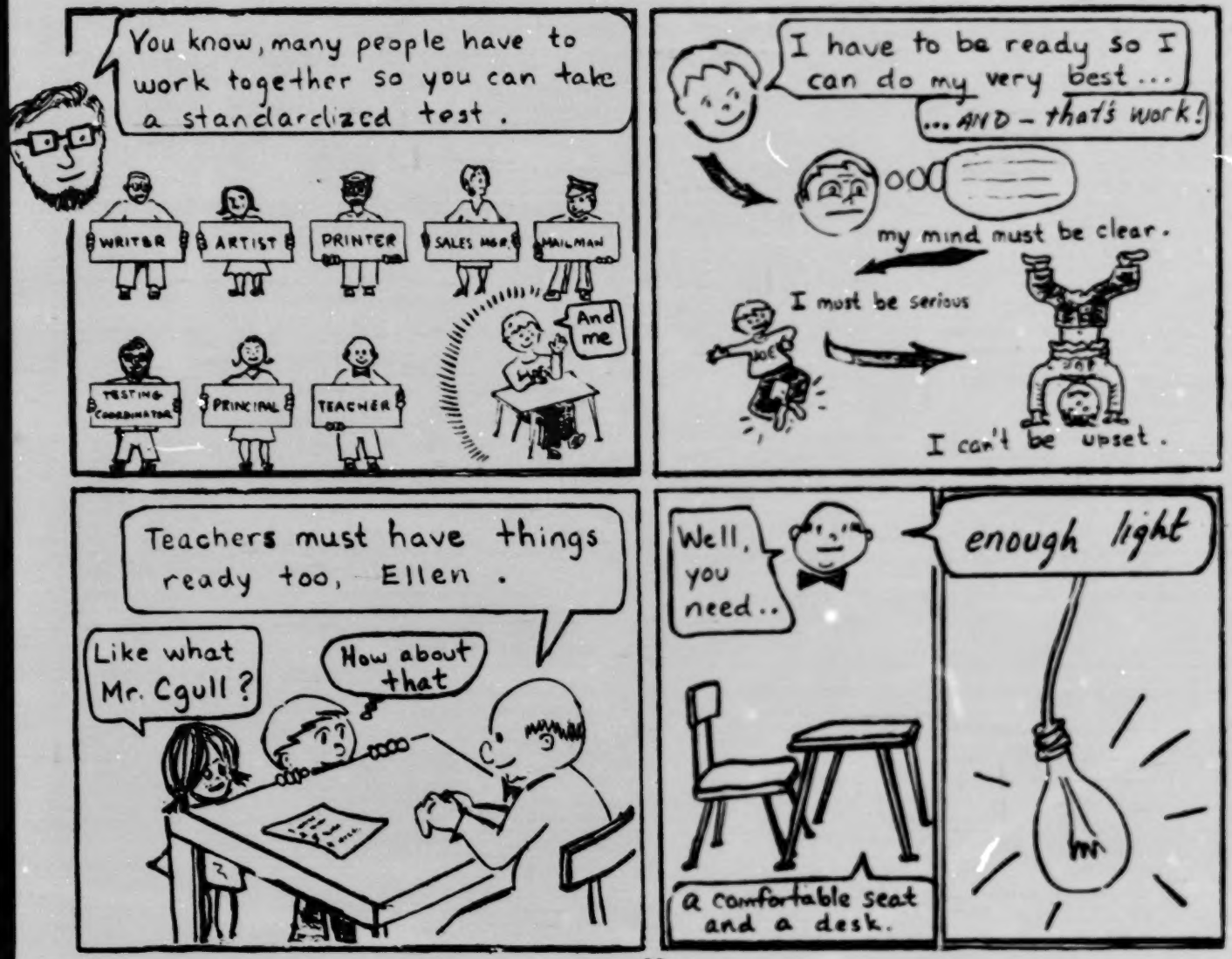
DURING (OR AT THE TIME OF) TESTING:

This section has been expanded into cartoon format since its content is of importance to you, the test administrator, and to the student. Reproductions of this section might be given to students to assist in their pretest preparation.

PROFESSOR
MINDE X. PANDOR
AND HIS FRIENDS
I HATE TESTS!
Why, Ellen?
Oh, it's always such a problem. I get nervous; I worry...
SIGH!
the night before a test I can't sleep.
Ooo

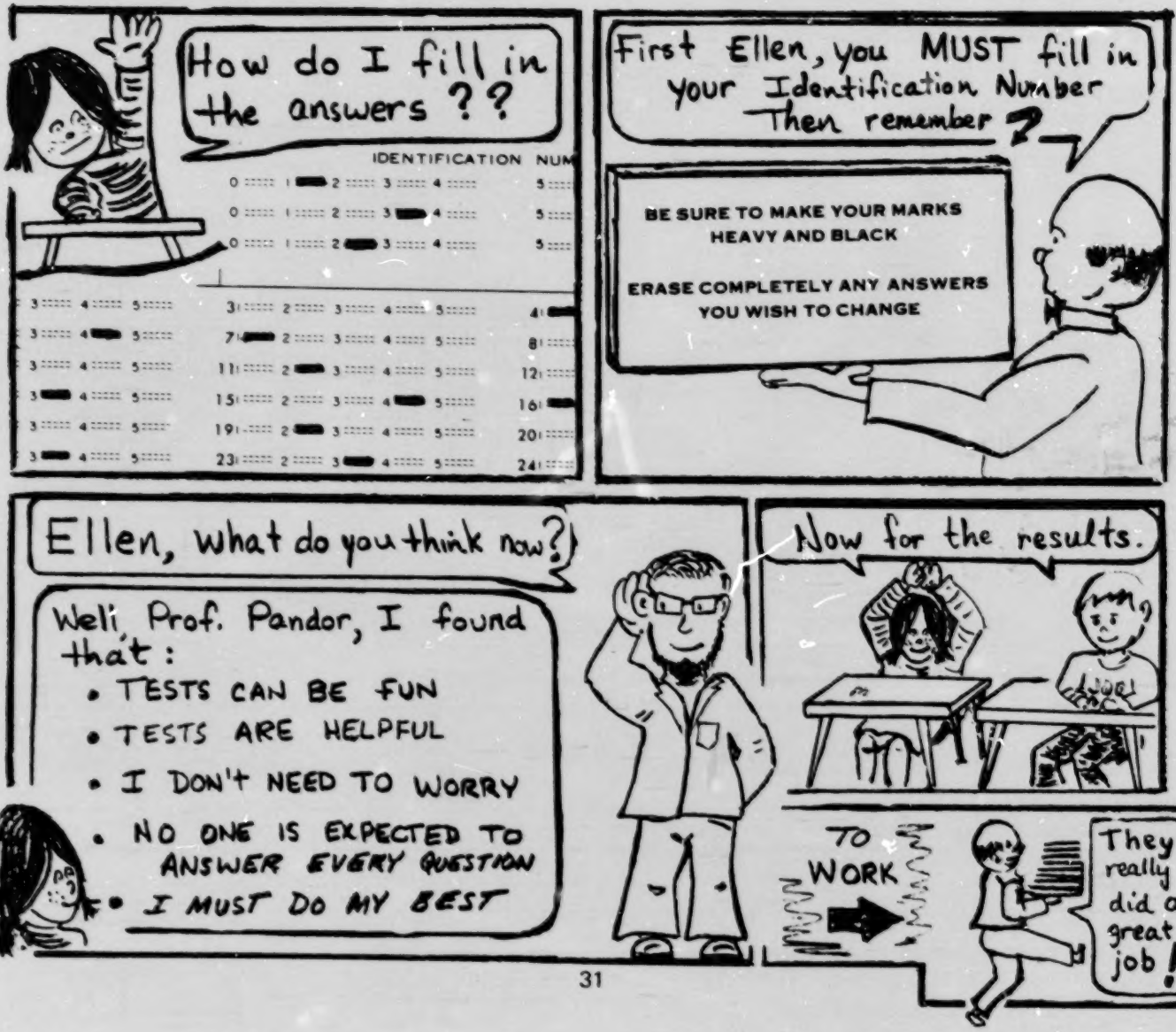
Gee, that's too bad, you shouldn't feel that way.....
I think tests are very helpful.
Really!
Sure, tests have a lot of interesting things to do.
So what! Who do tests help?
You. And our teacher, Mr. Cgull....
STANDARDIZED TESTS


That's right Joe, tests help me to see what you have learned...
Sooo...
.. I can help you on things you still need to learn.
I like tests because I can find out how much I can really do!
JOE
But some questions are so HARD!
A difficult question for you, Ellen might be easy for someone else.
PROF. PANDOR
If 4×3 = □ and □×6 = 3 then 4 + □ + (6×□ = 3) × 17/33 ÷ 3 = 50
Some questions are difficult for everybody
Nobody can answer all of the questions.
That makes me feel better.
JOE


You know, many people have to work together so you can take a standarclized test.
WRITER
ARTIST
PRINTER
SALES MGR.
MAILMAN
TESTING COORDINATOR
PRINCIPAL
TEACHER
And me
I have to be ready so I can do my very best...
...AND - that's work!
my mind must be clear.
I must be serious
I can't be upset.
Teachers must have things ready too, Ellen.
Like what Mr. Cgull?
How about that
Well, you need..
a comfortable seat and a desk.
enough light

and enough space
HELP!
there should be no distractions.
you need good fresh air.
So I don't fall asleep.
Is there anyone who doesn't have a pencil .... scratch paper
... answer sheet...
Test ? ... paper
READY!
Room 122
Mr. J.L Cgull
TESTING DO NOT DISTURB

WAIT
There is more for you to do too, Mr. Cgull.
YOU HAVE TO BE FAIR TO EVERY STUDENT
HOORAY!!
To be fair, everyone must be given exactly the same amount of time for each test.
Yes, and as I watch the time my students can follow the signs in the test book.
GO on to next page
STOP
STOP


How do I fill in the answers ??
IDENTIFICATION NUM
First Ellen, you MUST fill in your Identification Number Then remember ?
BE SURE TO MAKE YOUR MARKS HEAVY AND BLACK
ERASE COMPLETELY ANY ANSWERS YOU WISH TO CHANGE
Ellen, what do you think now?
Well Prof. Pandor, I found that:
• TESTS CAN BE FUN
• TESTS ARE HELPFUL
• I DON'T NEED TO WORRY
• NO ONE IS EXPECTED TO ANSWER EVERY QUESTION
• I MUST DO MY BEST
Now for the results.
TO WORK
They really did a great job!

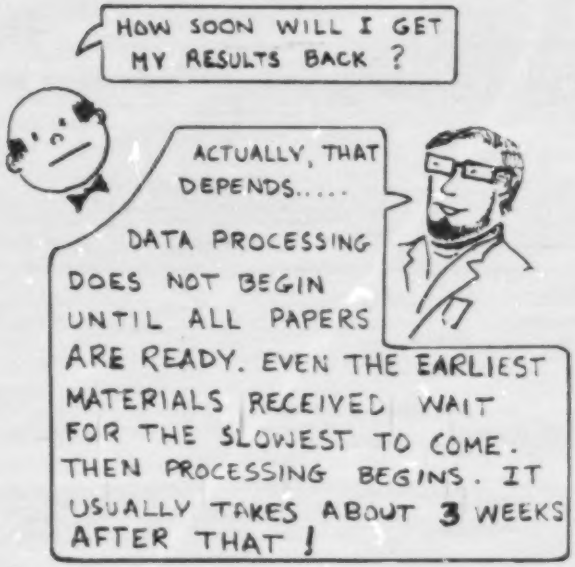
DURING (OR AT THE TIME OF) TESTING. [This section summarizes and expands the information in the cartoon section.]

1. Set the proper tone for a good test situation.
   a. Emphasize that students should do their best in answering each question.
   b. Explain to students about the test (such as, "Some of the questions will be difficult and students are not expected to know all of the correct answers.").
   c. Let students know some hints for effective test taking.
      1) Tell students to skip a really difficult question and go on to the next. The object is to show how many they can answer, so they should not get bogged down on one sticky question. Then, if there's time, they can go back and try the ones they skipped.
      2) Advise students not to go back and change answers. Their first thoughts are usually best.
      3) If guessing is penalized in the scoring, encourage students to answer judiciously. If they think they have any idea of the right answer, fine, but discourage random guessing.
      4) If an "I don't know" choice is available, emphasize that if truly accurate, this is a perfectly proper answer. (It should not be used as a cop-out, though.)
2. Set up a proper testing environment.
   a. Provide each student with adequate, comfortable seating space. Separate students to prevent distractions and copying.
   b. Have all desks or tables cleared of materials except for pencils and erasers.
   c. Post a sign (Testing - Do Not Disturb) to prevent disruptions while testing is underway.
3. Pass out supplementary materials. Make sure all students have enough pencils, scratch paper, if necessary, etc.
4. Follow the publisher's directions.
   a. The directions that accompany standardized tests are very detailed and specific. They should be followed exactly to assure uniformity of testing within the district and for comparability of results to the normative data.
   b. Explain procedures and directions to students such as:
      1) This is a timed test, do not open the test booklet until instructed to do so.
      2) There is no passing or failing to the test, do your best.
      3) All answers must be put in the appropriate spot on the answer sheet (if separate sheet is used).
      4) At this time we will complete only the reading (for example) section of the test.
      5) Be careful not to make stray marks on your answer sheet.
      6) Here are your test papers, let's do the practice question, etc., etc.
   c. Ask if there is anyone who does not understand what is to be done.
      Clarify any directions that may still be unclear.
   d. Begin the test at a known time; monitor students during the testing; make sure all

students stop work on the test exactly in accordance with the specified time limits. Don't be sympathetic and give students an extra minute!

Time limitations are not punishment techniques, they are to assure everyone equal opportunity. Time limitation is the way of controlling one of the variables that can affect the ability to compare results in a valid manner.

5. If more than one section of a test is being administered, allow for a brief rest period between work periods.


HOW SOON WILL I GET MY RESULTS BACK ?
ACTUALLY, THAT DEPENDS.....
DATA PROCESSING DOES NOT BEGIN UNTIL ALL PAPERS ARE READY. EVEN THE EARLIEST MATERIALS RECEIVED WAIT FOR THE SLOWEST TO COME. THEN PROCESSING BEGINS. IT USUALLY TAKES ABOUT 3 WEEKS AFTER THAT !

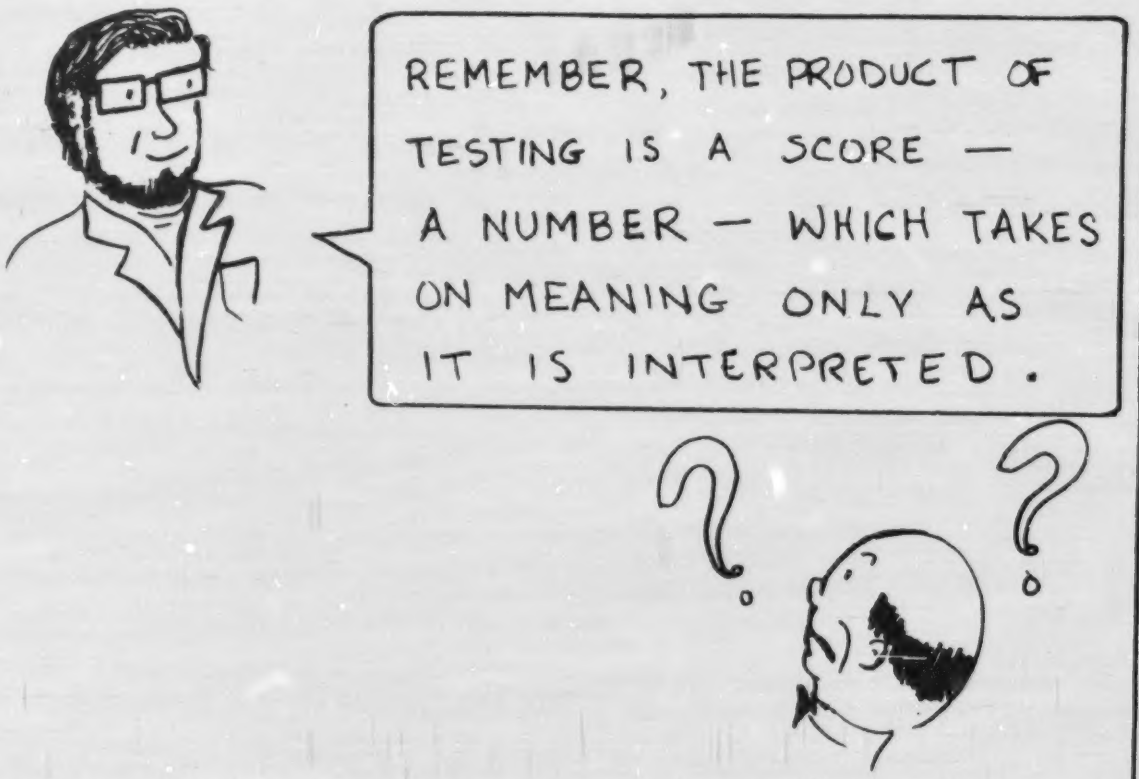
AFTER TESTING

1. Collect and account for all testing materials, keeping test booklets separated from answer sheets.
2. Scan all answer sheets to assure that each student's name and identification number have been filled in correctly.
3. Organize materials in the way you want to have the testing results organized. Whether tests are to be hand scored or machine scored, organization saves wasted time, confusion and potential error. If test results are desired by individual class, papers should be grouped this way. If results are desired on a grade level basis for an entire school, they should be so grouped.
4. Send all answer sheets immediately to the Research and Evaluation Office so scoring and data processing can begin, excepting hand score materials to be done by the tester. (However, since testing situations may vary from time to time, follow the specific directions relative to each testing situation.)
5. Prepare yourself and the students for the return of information while scoring and data processing are taking place. What will results mean to you? How will the information be used?

Since the value in doing all of the foregoing procedures lies in using the information obtained from testing, Interpretation and Uses of Data is the subject of the entire next section of this handbook.

■■■

INTERPRETATION AND USES OF DATA



In this section, a perspective on interpretation and uses of standardized test data is presented by discussions about:

PURPOSE
NORMS
TEST SCORES
TYPES OF REPORT FORMATS
USES OF INDIVIDUAL STUDENT DATA
USES OF GROUP DATA

In order to use standardized test data appropriately, and not misuse such data, one needs to understand considerable background information. This includes knowing the purpose(s) for standardized testing, how standardized tests are made, how they differ from other types of tests, their strengths and their limitations. Knowing all this, subjective judgment must bring compassion and morality to the use of data, remembering that testing is only a tool, not an end.

INTERPRETATION AND USES OF DATA

PURPOSE

The following is an excerpt from "Ralph Tyler Discusses Behavioral Objectives," Today's Education, Sept/Oct 1973:

> The basic purpose of (standardized) testing is to take a total group and arrange (the members) in some kind of order so that you can say here is the top 10 percent and here is the bottom 10 percent. The population is arranged on a linear scale from the best to the worst. This is called norm referenced testing.
>
> When this type of test is being made, various test items are tried out. If the items differentiate among the persons tested, they are retained. So a typical achievement test has about 80 percent of its items in the narrow range of difficulty where between 40 and 60 percent of people tested get the answer right.
>
> If the purpose is to identify those who do best on the total test and those who do poorest, this is an efficient way to go about it. But if you trying to answer the question, "What have students learned?" you run into difficulties. This is because (a) almost all items that most persons can answer correctly are dropped from the typical achievement test because they do not discriminate and (b) those items that almost no one can answer correctly are also dropped because they don't discriminate either.
>
> Actually, instead of testing what our students have learned, we have been using test items to differentiate some students from other students.

This purpose behind standardized testing accounts for one of the major limitations on use of the information derived from **standardized tests**. Many people are tempted at this point to throw out standardized tests because they don't do everything that is needed. In this situation it is vital to understand the strengths and uses of standardized test data so one does not "throw out the baby with the bath water." Various aspects of standardized tests should be understood in order to put strengths and weaknesses into functional perspective. Thus, the next three sections examine characteristics related to standardized tests.

NORMS

The word **norm**, from normal, can be thought of as a synonym for average. Norm referencing simply means that **achievement** is related to how others do on the same test.

Norms are descriptions of the achievement of typical students in some large population, not goals nor standards that set expectations. The foregoing point (repeated from page 17) cannot be overemphasized, for there is commonly misunderstanding about the use of standardized test normative data as goal standards. Fairly subtle differences distinguish appropriate use from misuse. While norms are not standards that set expectations, they may none the less be valid comparative values against which to reflect group achievement data.

For a district to accept normative averages as indices for comparison of growth values from year to year may be quite appropriate particularlt where characterististics of

the population from which the norms were taken are similar to the local population. Using norms (averages) as standards for individual students, however, is a dangerous misuse.

Students of all capabilities contribute their achievement scores to create the average figure that is the group norm. It is entirely natural for an individual to be below the norm just as it is natural for another individual to be above the norm. Without such cases, the norm would not be what it is! Expecting an individual to meet normative figures would require some students to stretch beyond what is natural for them, others to achieve less than their natural level. Thus norms can be appropriately stated in the district's policy statement as standards for comparison, yet be most inappropriate as goals or standards for individuals.

TEST SCORES

The product of testing is a score. It is important to remember two things about test scores:

1. All of the correct and incorrect responses that a student makes on a test (or subtest) are condensed to yield the single numerical score. While this decreases the information to a manageable level, it concomitantly loses the range of detail in the information.
2. Any score is only a estimate. If tested again and again, with alternate forms and on other days, the student's scores would be expected to vary somewhat; any one score would be an approximation or estimate of an ideal true score. This relates to the concept of **error of measurement.**

Scores from tests can be reported in any of several forms. The **raw score** simply tells how many items the student answered correctly. From this basic score, all other converted scores derive. The various ways that converted scores may be expressed, such as percentages, **standard scores, percentiles, stanines, grade equivalents,** have different characteristics and applications. Brief descriptions of these terms are found in the glossary; the following statements give some perspectives on the nature of each of the various scores.

Raw Scores

Due to differences between tests, the number of items they contain and the difficulty of items, it is not appropriate to directly compare raw scores from one test to raw scores from another test. For many statistical purposes however, the raw score is the most important and valid value to use.

Percentages

Probably the most widely used converted scores are percentages. These simply adjust all raw scores to a common base as if every test had 100 items. The percentage is then the statement of how many of the 100 items a person would have scored correctly.

As with raw scores or standard scores, one must be cautious about comparing scores of different tests. This is commonly abused, as when people say, "You did better on the math test (80%) than you did on the spelling test (75%)." There is no proper basis for such a statement, for quite possibly the spelling test was comparatively much more difficult than was the math test.

Standard Scores

A scale that is created to adjust raw scores into a more useful, relatable form is made up of standard scores. There

may be any number of different standard score scales. Most standard scores relate only to the specific test for which they were derived, such as the standard score scale for the MAT reading sub-section. Thus a standard score of 58 has meaning only as one understands the particular scale of which it is a part.

Percentiles

Once a person's score is determined, it may be related to how well other people did on the same test. Percentiles are ways of expressing relative performances among individuals or groups. If any hundred people are tested someone must have the lowest score, someone the highest, someone would be in the 60th position. A score coinciding with the 60th percentile is regarded as equaling or surpassing that of 60 percent of the persons in the group, with 40 percent of the performances exceeding this score. A percentile value tells nothing about how many items a person scored correctly, it simply expresses how well one person did in relation to others.

It is also important to know that the percentile scale is not made up of equal units. Thus, there is no absolute relationship between percentile scores; the actual difference in achievement between persons in the 20th and 30th percentile ranks may be considerably different than the difference in achievement between persons scoring in the 50th and 60th percentile ranks. Because percentile units are not equal, they cannot be added, subtracted, averaged, etc.

The usefulness of percentiles to the teacher is limited to giving an impression of how well a student achieved on a test compared to others. By looking at the percentile ranking of a student on several tests, one gains a perception of a student's general capability relative to others.

Stanines

A stanine is one of the steps in a nine-point scale of standard scores. The stanine (short for standard-nine) scale ranges from a high of 9 to a low of 1, the mean being 5. (See glossary for diagram.)

Stanines are increasingly popular since they are broad categories and thus prevent some of the overprecision associated with the way in which people interpret percentiles, IQ scores, etc. This is a real advantage. That stanines are wide range scores frustrates some people who want to be more exact, but that is usually a good thing. Few tests are capable of measuring as exactly as people wish and attempts at pinpoint accuracy are generally examples of data misuse.

Tests and their results, as most people know, are subject to error and limitations; that is, any score is an approximation rather than an exact, absolutely accurate measurement value. Stanines, being rather broad, gross measures are a range within which finer distinction is not possible. It is better to see a student's position confidently within a general category than to be unwarrantedly concerned about minor differences on a more precise scale. For example, consider two students whose scores were stated in percentile ranks of 43rd percentile and 57th percentile. It is compelling to want to assume a fairly big difference between these two, yet both would be expressed as stanine 5, with no significant instructional difference being certain.

Another advantage to stanine usage is the relatability of standings. Stanines obtained for a distribution of scores or similar data are comparable with any other set of stanines obtained for the same group of individuals.

Stanines also permit statistical manipulation without serious distortions.

Grade Equivalents

These scores have been used for decades and according to most people are "the ones I can understand." Unfortunately, along with being simple and easy to understand (superficially), they are the most easily misinterpreted score and misuse is common. Although G.E.'s are widely used by popular demand, they are basically discredited by professionals in measurement.

A grade equivalent indicates the grade placement of pupils for whom a given score is average. For example, if a raw score of 37 on a math test corresponds to a grade equivalent of 5.2, it means that 37 is the typical (average) grade for students in the second month of grade 5. It is most important to emphasize that this is average performance (a statistical concept below which half of a population would exist, by definition). Grade equivalents are not directly comparable for students who are above or below average in performance.

Another reason why G.E.'s are losing popularity is that they are not equal throughout the scale. An increase of 8 raw score points at one point in the scale might show as three month of growth while in another part of the scale this would be two years growth. Laymen frequently think that a student should be in the grade corresponding to the G.E. score. Actually a student in grade 4 could receive a G.E. of 6.7 on a math test but have no ability to work with the math skills in decimals and fractions in the grade 6 curriculum. Such a G.E. score simply means that the student is performing on grade 4 work well above the average fourth-grader.

Other limitations make G.E. scores objectionable. These include: the assumption that a 12 month gain really occurs in the 9 or 10 month school year; the unwarranted extrapolation to levels not covered by a given test; the assumption that growth is regular and equal on a month to month basis; and mathematical functions cannot be performed on G.E. scores.

Probably the most dangerous aspect of G.E.'s is that they are frequently considered as standards of performance. This is inappropriate for it would cause everyone to be expected to be average. The inappropriateness problem of being considered as a performance standard of expectation is true of all scores, but it is much more commonly abused with grade equivalents simply because of the way they are stated. It seems reasonable to think that grade level, or grade equivalent, should be an appropriate expectation until one realizes that grade level is a statement of average performance.

TYPES OF REPORT FORMATS

After test papers are completed by students, organized, and sent to data processing where scoring services are performed, how are results returned? The formats from various scoring services and for different tests may vary from one another, but basic types of information are available. These include **class lists**, **frequency distributions**, statistical data, item analysis, and some other more specialized forms.

Class Lists

The class list (list report of pupil scores) provides the name of each student with scores expressed in score types (raw scores, stanines, percentiles) as requested of the scoring service. The list is accompanied by a duplicate listing on gummed labels so the information can be posted directly onto the permanent record. Figures 5.1 and 5.2 show the formats as they are currently obtained through the Board of Cooperative Educational Services (BOCES) and Harcourt Brace Jovanovich. Regardless of the specific format, similar information is presented on these forms.

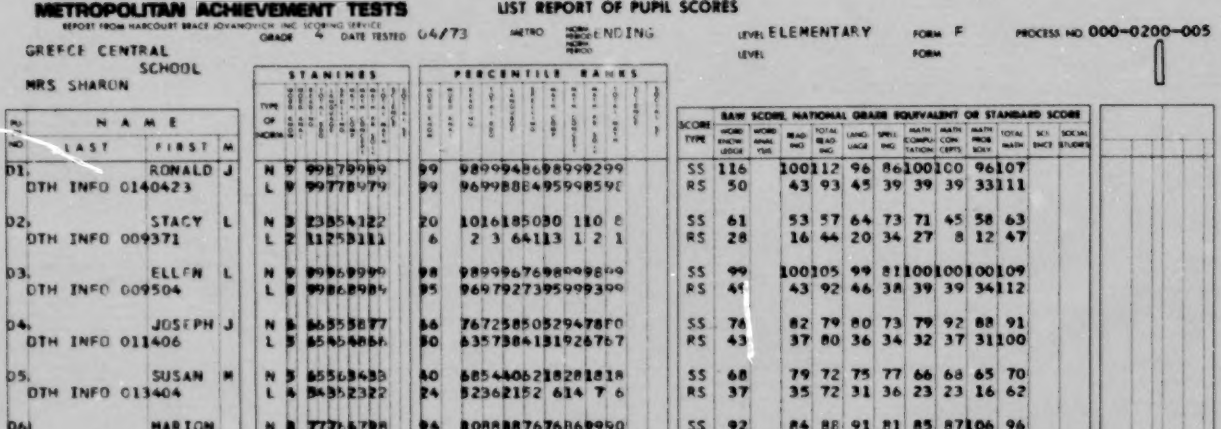
```
043310          NANCY              *METROPOLITAN ACH* GR. 05 05/73 INT FORM F
TEST-WD KNG TOT.R RDNG SPELL LANG COMPU CONCPT TOT.M PROB SOC.ST SCI
S/S -  90    90   88    95   102    89    90    94   93
G.E.- 7.9   7.6  7.3   8.3  9.8   6.3   7.3   6.7  7.1     .      .
%-S9- 82 7  82 7 80 7  90 8 92 8  66 6  74 6  74 6 80 7

045244            JOSEPH          *METROPOLITAN ACH* GR. 05 05/73 INT FORM F
TEST-WD KNG TOT.R RDNG SPELL LANG COMPU CONCPT TOT.M PROB SOC.ST SCI
S/S -  87    82   76    82   82    83    80    86   83
G.E.- 7.3   6.2  5.3   6.1  5.7   5.6   5.5   5.4  5.6     .      .
%-S9- 77 7  62 6 40 5  64 6 44 5  42 5  46 5  46 5 52 5

079622            TRACEY          *METROPOLITAN ACH* GR. 05 05/73 INT FORM F
TEST-WD KNG TOT.R RDNG SPELL LANG COMPU CONCPT TOT.M PROB SOC.ST SCI
S/S -  66    65   67    70   70    87    73    83   70
G.E.- 3.9   3.9  4.1   4.4  4.0   6.1   4.6   5.1  3.9     .      .
%-S9- 16 3  16 3 18 3  24 4 16 3  62 6  30 4  36 4 18 3

100475          DOUGLAS            *METROPOLITAN ACH* GR. 05 05/73 INT FORM F
TEST-WD KNG TOT.R RDNG SPELL LANG COMPU CONCPT TOT.M PROB SOC.ST SCI
```

Figure 5.1



METROPOLITAN ACHIEVEMENT TESTS — LIST REPORT OF PUPIL SCORES

REPORT FROM HARCOURT BRACE JOVANOVICH, INC. SCORING SERVICE

GRADE 4 DATE TESTED 04/73 METRO NORM PERIOD ENDING LEVEL ELEMENTARY FORM F PROCESS NO 000-0200-005

GREECE CENTRAL SCHOOL

MRS SHARON

NO.	NAME LAST	NAME FIRST	M	TYPE OF NORM	STANINES	PERCENTILE RANKS	SCORE TYPE	RAW SCORE: NATIONAL GRADE EQUIVALENT OR STANDARD SCORE: WORD KNOWLEDGE	WORD ANALYSIS	READING	TOTAL READING	LANGUAGE	SPELLING	MATH COMPUTATION	MATH CONCEPTS	MATH PROB SOLV	TOTAL MATH	SCIENCE	SOCIAL STUDIES
01		RONALD	J	N	9	99	SS	116		100	112	96	86	100	100	96	107		
	DTH INFO 0140423			L	9	99	RS	50		43	93	45	39	39	39	33	111		
02		STACY	L	N	3	20	SS	61		53	57	64	73	71	45	58	63		
	DTH INFO 009371			L	2	6	RS	28		16	44	20	34	27	8	12	47		
03		ELLEN	L	N	9	98	SS	99		100	105	99	81	100	100	100	109		
	DTH INFO 009504			L	8	95	RS	49		43	92	46	38	39	39	34	112		
04		JOSEPH	J	N	6	66	SS	76		82	79	80	73	79	92	88	91		
	DTH INFO 011406			L	5	50	RS	43		37	80	36	34	32	37	31	100		
05		SUSAN	M	N	5	40	SS	68		79	72	75	77	66	68	65	70		
	DTH INFO 013404			L	4	24	RS	37		35	72	31	36	23	23	16	62		
06		MARION		N	8	94	SS	92		84	88	91	81	85	87	106	96		
	DTH INFO 109298			L	8	89	RS	48		38	86	43	38	35	35	35	105		
07		MARY		N	7	84	SS	83		82	83	89	77	92	80	83	89		
	DTH INFO 029818			L	6	73	RS	46		37	83	42	36	37	31	29	97		
08		DONNA	E	N	3	14	SS	58		62	58	67	77	74	66	67	72		
	DTH INFO 038912			L	1	3	RS	23		23	46	23	36	29	21	18	68		
09		TERRY	M	N	8	94	SS	92		77	82	86	75	95	87	96	97		
	DTH INFO 068424			L	8	89	RS	48		34	82	40	35	38	35	33	106		
10		JERRY	J	N	5	60	SS	68		72	69	77	70	80	61	70	74		
	DTH INFO 082850			L	6	24	RS	37		31	68	33	31	33	18	20	71		
11		SUZANN		N	6	82	SS	74		80	77	88	66	88	73	76	83		
	DTH INFO 083354			L	5	64	RS	42		36	78	41	39	36	27	25	88		
12		MARYAN		N	8	90	SS	87		89	90	99	94	100	90	106	105		
	DTH INFO 083484	8		L	7	81	RS	47		40	87	46	40	39	36	35	110		
13		MICHAE	E	N	2	8	SS	54		56	53	61	65	80	64	76	78		
	DTH INFO 163760			L	1	1	RS	18		18	36	18	26	33	20	25	78		

(See notes on back)

Figure 5.2

Frequency Distributions and Statistical Data

In order to get an overview of group results, data are summarized as descriptive statistics. **Means, medians, quartiles** and distributions of scores over the potential range are presented for each test or subtest taken. Such data are compiled for each group as submitted to the scoring service. As with other types of information forms, specific formats may vary. See Figures 5.3 and 5.4.

GREECE
CENTRAL SCHOOL DISTRICT

BUILDING NAME

STATISTICAL DATA

METRO 10
TEST NAME

WRD KNOW STATISTICS
SUBTEST

FI FORM | 05 GR | 05 MO | 72 YR

PAGE 1
H. R. NO. OR TEACHER NAME

S/S

SCORE GROUP	CUM FREQ	f	LOCAL NORMS
2	1	1	1
3	1		1
4	1		1
5	1		1
6	1		1
7	2	1	3
8	3	1	4
9	3		4
10	3		4
11	3		4
12	4	1	5
13	6		5
14	6	2	8
15	15		
18	18	3	23
29	21	3	26
30	23	2	29
31	23		29
32	26	3	33
33	29	3	36
34	30	1	38
35	39	9	49
36	46	7	58
37	49	3	61
38	52	3	65
39	57	5	71
40	61	4	76
41	63	2	79
42	66	3	83
43	70	4	88
44	75	5	94
45	78	3	98
46	78		98
47	80	2	99+

	RAW SCORE	~~GR EQUIV~~	%ILE	STA 9
MEAN	33.65	82	66	6
MEDIAN	35.64	84	70	6
Q_1	29.16	77	50	5
Q_2	35.64	84	70	6
Q_3	40.25	90	82	7

STANDARD DEVIATION 9.62

Figure 5.3

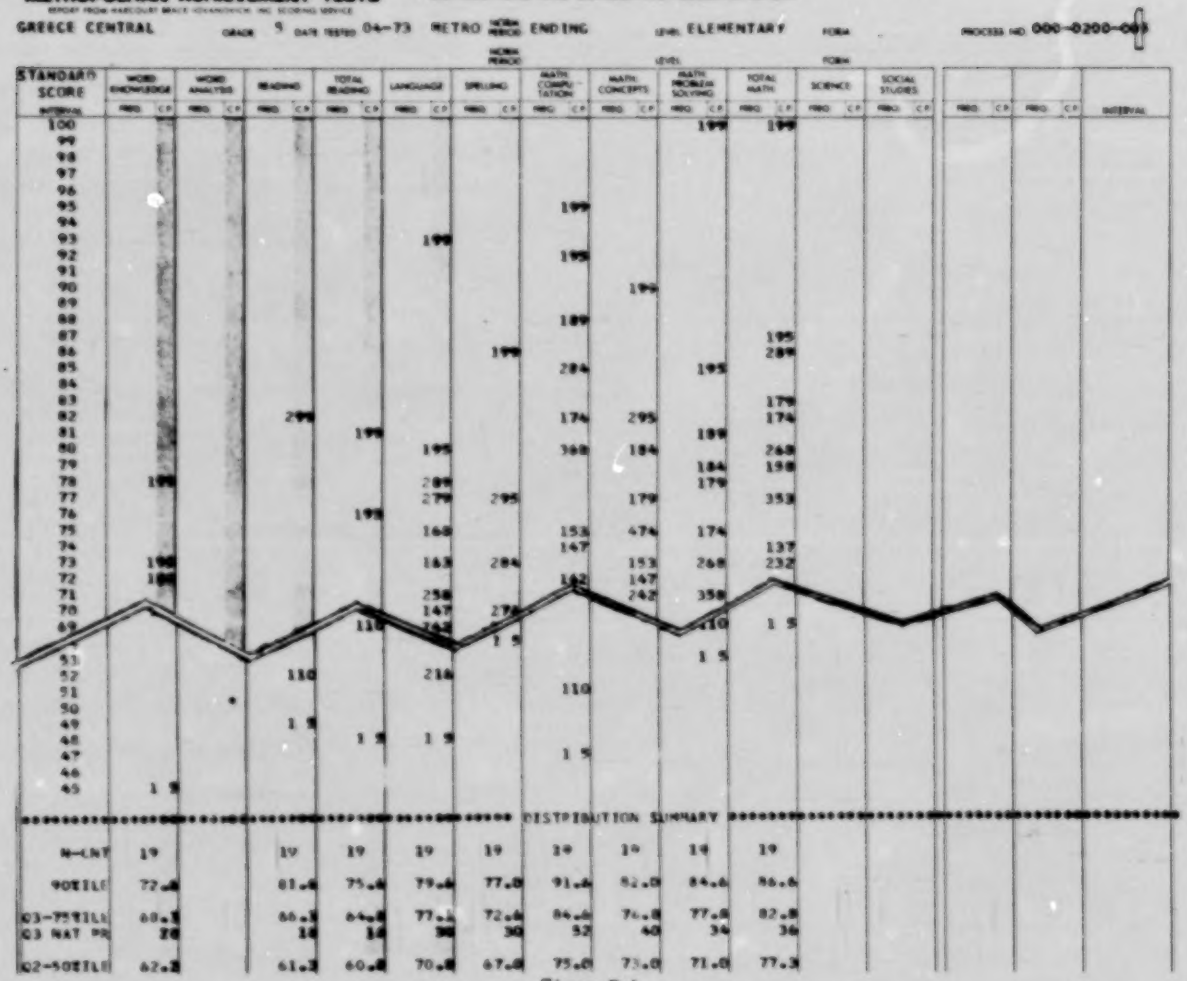
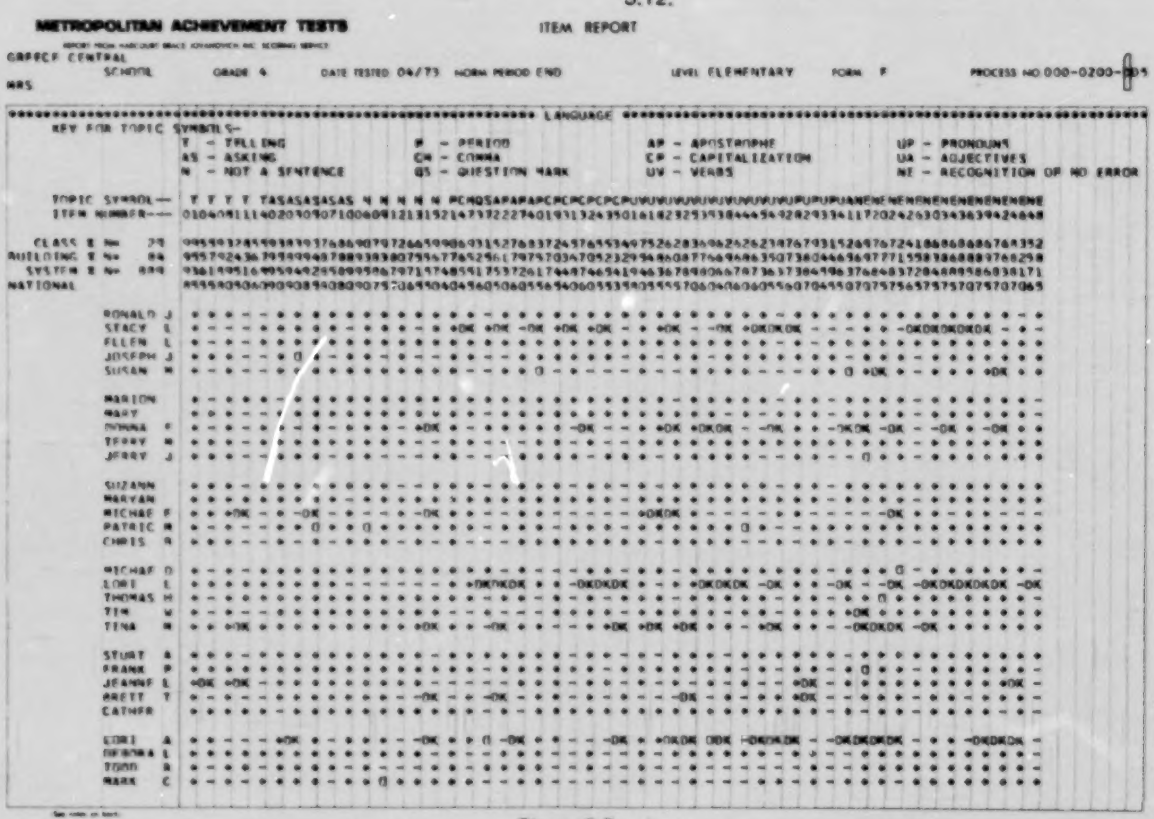

METROPOLITAN ACHIEVEMENT TESTS
DISTRIBUTIONS AND CUMULATIVE PERCENTS REPORT
GREECE CENTRAL
ELEMENTARY
STANDARD SCORE
WORD KNOWLEDGE
WORD ANALYSIS
READING
TOTAL READING
LANGUAGE
SPELLING
MATH COMPUTATION
MATH CONCEPTS
MATH PROBLEM SOLVING
TOTAL MATH
SCIENCE
SOCIAL STUDIES
FREQ.
C.P.
INTERVAL
DISTRIBUTION SUMMARY
N-CNT
90%ILE
Q3-75%ILE
Q3 NAT PR
Q2-50%ILE

Figure 5.4

Item Analysis

How many students answered question 17 correctly? An answer to such a question can be helpful to the teacher or curriculum specialist for purposes of program evaluation by examining results on a group basis. Using item analysis data is described in greater detail in the section on Uses of Group Data later in this chapter and printout formats are shown in Figures 5.5, 5.6, 5.9 and 5.12.



METROPOLITAN ACHIEVEMENT TESTS ITEM REPORT

GRPRCP CENTRAL
SCHOOL GRADE 4 DATE TESTED 04/73 NORM PERIOD END LEVEL ELEMENTARY FORM F PROCESS NO 000-0200-005
MRS

LANGUAGE

KEY FOR TOPIC SYMBOLS—
T - TELLING P - PERIOD AP - APOSTROPHE UP - PRONOUN
AS - ASKING CM - COMMA CP - CAPITALIZATION UA - ADJECTIVES
N - NOT A SENTENCE QS - QUESTION MARK UV - VERBS NE - RECOGNITION OF NO ERROR

TOPIC SYMBOL— T T T T TASASASASAS N N N N N PCMQSAPAPAPCPCPCPCPUVUVUVUVUVUVUVUVUVUPUPUPUPUNENENENENENENENENE
ITEM NUMBER— 0104091111402030907100609121315214737222740193132435016182925935844454928293341172024263034363942464848

CLASS % No 78 9959932855909839376869079726659906893155276837245765534975262836962626231876793152697672418868688676843S2
BUILDING % No 84 9955702436799499687889838075967765256179757037052329548608779699863507380446969777155838688876682S8
SYSTEM % No 889 9361499516995949259099586279715748554175372617446746541946367898067673633049863726648372044895860381T1
NATIONAL 8999490906399038990880907S7086490404960508655654606055539405555706060606056070455079757565757570757070S5

RONALD J
STACY L
ELLEN L
JOSEPH J
SUSAN M

MARION
MARY
THOMAS P
TERRY M
JERRY J

SUZANN
MARYAN
MICHAE P
PATRIC M
CHRIS R

MICHAE D
LORI L
THOMAS H
TIM W
TINA M

STUAT A
FRANK P
JEANNE L
BRETT T
CATHER

LORI A
DEBORA L
TODD B
MARK C

Figure 5.5

GREECE METROPOLITAN ACHIEVEMENT - INTERMEDIATE LEVEL

ITEM ANALYSIS

READING — CRAIG HILL — GRADE 5

QUESTION NO.	NUMBER OF STUDENTS RESPONDING						PERCENTAGE RESPONSE					
	NO RESPONSE	RESPONSE NUMBER 1	2	3	4	5	NO RESPONSE	RESPONSE NUMBER 1	2	3	4	5
1		4		2	62*		.0	5.9	.0	2.9	91.2*	.0
2	1	1	3	59*	4		1.5	1.5	4.4	86.8*	5.9	.0
3			50*				.0				2.9	
4		10			50*			14.7	2.9	8.[illegible]		[illegible]
9		41*	2	12	13		.0	60.3*	2.9	17.6	19.1	.0
10		57*	5		6		.0	83.8*	7.4	.0	8.8	.0
11		5	3	53*	7		.0	7.4	4.4	77.9*	10.3	.0
12	2	3	31*	14	18		2.9	4.4	45.6*	20.6	26.5	.0
13		47*	3	5	13		.0	69.1*	4.4	7.4	19.1	.0
14		1	1	3	63*		.0	1.5	1.5	4.4	92.6*	.0
15		62*	4	1	1		.0	91.2*	5.9	1.5	1.5	.0
16			2	63*			.0			92.6*	.0	.0
	[illegible]0	23*			17			33.8*	13.2	[illegible]		
41	15	5	16	19*	13		22.1	7.4	23.5	27.9*	19.1	.0
42	14	7	31*	10	6		20.6	10.3	45.6*	14.7	8.8	.0
43	13	22	10*	16	7		19.1	32.4	14.7*	23.5	10.3	.0
44	13	39*	5	7	4		19.1	57.4*	7.4	10.3	5.9	.0
45	12		9	27*	20		17.6	.0	13.2	39.7*	29.4	.0

* DENOTES RIGHT RESPONSE

Figure 5.6

Less Commonly Used Forms

Rank Listing are simply class lists that instead of being arranged alphabetically are ordered from highest to lowest achievement for some test or subtest.

The *Item Report* form from Harcourt Brace Jovanovich for the MAT is shown in Figure 5.5. This specific type of item analysis is really two separate reports (the Basic Item Report and the Pupil Item Report) combined on the same sheet since they are closely related. Pupils' responses are shown for each test item through use of symbols that indicate correct, wrong, don't know and did not answer. These data are helpful for looking across a pupil's report for areas of weakness and strength.

The upper portion of the Item Report consists of group item analysis information. While this derives from the compilation of individual pupil responses, the percentages shown for each test item lose pupil identity. The use of item analysis data (described under Uses of Group Data) is for program evaluation. BOCES data processing facilities at present do not offer pupil item reports.

Class Analysis Charts provide a grouping of students into stanine categories. This type of printout from Harcourt Brace Jovanovich essentially does the same thing as a rank listing by each subtest. Such categorization of students by stanine (or some other criterion) can be done easily by hand also. The limited usefulness of such information, however, often discourages expending effort and/or expense of doing this analysis.

The Class Analysis Chart shows data on a one dimensional basis. When two criteria are used and individual cases are plotted into cells of a grid according to the scores in both of the criteria, the chart is said to be bivariate.

Bivariate Charts are not commonly used as part of our program. The chart pictured here (Figure 5.7) is to serve as information about this kind of capability. Such charts can be constructed by hand for any pair of criteria for which one wants to see how the members of a group distribute. For example, each student's report card grade in math and MAT math score can be plotted to see how well such values correlate for the group as a whole; scores on a teacher made test could be plotted against the scores from a Regents exam, etc.

TEACHER GRADE IN MATH

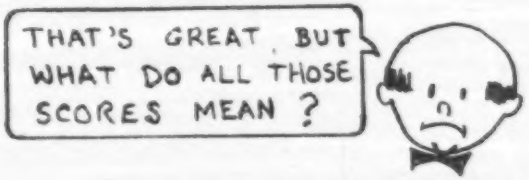
MAT MATH STANINE	F	D	C	B	A	TOTAL
9					//	2
8		/		////	///	8
7		/	/	///	//	7
6			//	////	//	8
5			~~////~~	///		9
4	/	/	///	/	/	7
3		///	///			6
2	/	//				3
1	//	/				3
Total	4	9	15	15	10	53

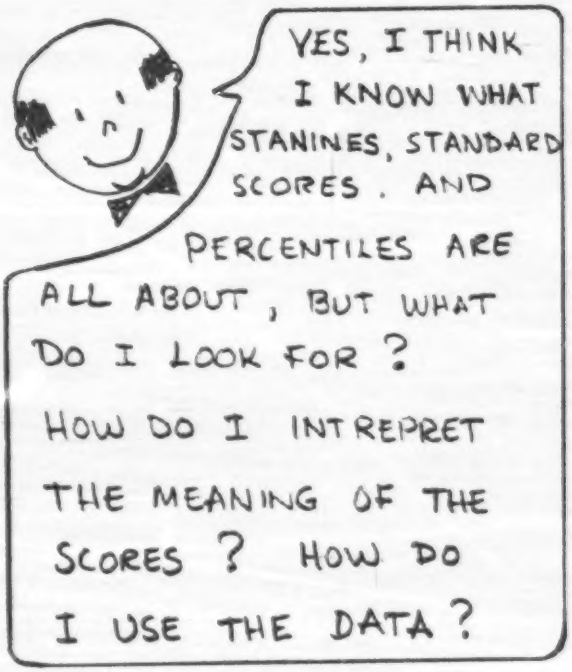
Figure 5.7

USES OF INDIVIDUAL STUDENT DATA

Data from standardized testing are useful in two general ways--for looking at individual students and for taking results collectively as group data. Uses of group data are described in the next section.

It is widely known that standardized test data can be misused and are thus looked poorly upon by some people. Abuses and misuses that do occur, however, do not diminish the value of test data, they simply amplify the need for knowledgeable application of such data. It is of primary importance to recognize that all scores are only estimates of theoretical true scores. A single test score may be affected by many factors: student ability and knowledge base, the student's health, interests, interpretation of questions, distractions during testing, etc. Keeping this tentativeness in mind, one must still assume that the actual score is a fair and accurate measure approximating the true score - - unless other information leads to another conclusion.




YES, I THINK I KNOW WHAT STANINES, STANDARD SCORES, AND PERCENTILES ARE ALL ABOUT, BUT WHAT DO I LOOK FOR ?
HOW DO I INTREPRET THE MEANING OF THE SCORES ? HOW DO I USE THE DATA ?

What to look for

Look for the usual. Look for each student's strongest and weakest subject areas across subtests. Compare results with other data that are available on each student to see if results on the present test are consistent with other information. A student who scores in stanine 2 in math, reading and spelling and in stanine 3 in language,

punctuation, and vocabulary presents a picture of low performance relative to others who comprise the "norm" group. Likewise a student with scores in stanines 7 and 8 is above average.

Looking for the usual or commonalities is often considered as not too important; there seems to be a much greater appeal or fascination with seeking the unique - - that which stands out. Though less exciting, common features are generally the base upon which decisions and programs should be built.

Look for the unusual. Inconsistency from one testing to another on the same content or across subtest areas raises the question of "why?". A discrepant area in a student's skill development provides a specific focal point for learning/teaching activities.

Don't overemphasize small differences. To be significant, differences between scores should be at least two stanines. For example, if a student's scores are as follows:

Area	Stanine
Reading	7
Language	6
Spelling	4
Math Concepts	6
Math Computation	7
Math Problem Solving	7

the only area that can be said to differ from the others is spelling.

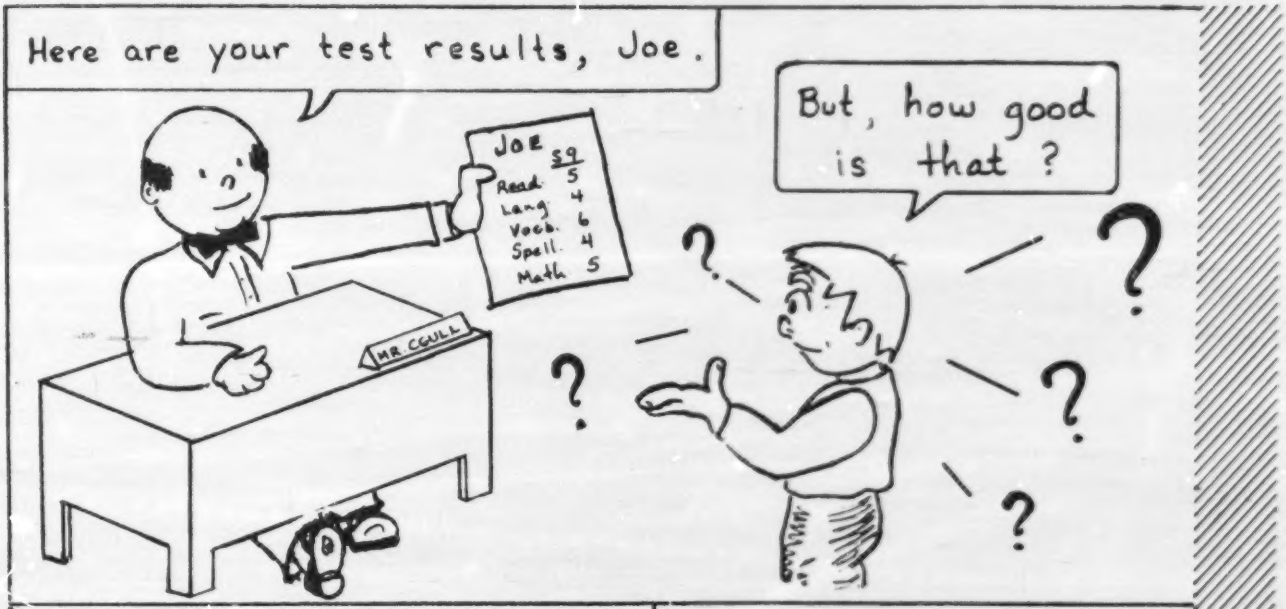
Don't expect to discover something new or different about every student. Oftentimes the results from a standardized test will only confirm the teacher's perception and understanding of a student.

Using Student's Scores

Instructional Applications. Information from standardized testing does not tell the teacher what is needed in order to carry out the next lesson, it simply gives perspective on relative proficiency. Expectations of more from standardized testing are unrealistically beyond what standardized tests are designed to provide and will lead to frustration and dissatisfaction.

Looking at the stanine performance of a student across several subtest skill areas allows identification of general strengths and weaknesses. For further information, more precise, diagnostic "testing" must be done. This need not be a paper/pencil test; questioning by the teacher with instant assessment of specific problems may very rapidly set direction for the next learning sequence. Alternately, if more "objective" data are desired, an assessment type test on specific objectives should be used for diagnostic purposes. Since only a small number of questions on a standardized test relate to a single skill, it is of tenuous validity to draw anything other than general conclusions.

Bases for selection. In addition to instructional applications, standardized test results are often used as selection criteria for scholarships, college admission, special programs, etc. Since the essence of standardized testing is to distribute people over a performance range, total scores over various subtests provide a fairly accurate relative comparison among those taking the test. Students who score highly on the academic skills tested are generally those who do well in future academic work. Thus, standardized tests such as Scholastic Aptitude Tests, Regents Scholarship Exams, etc. can be validly used to identify future success potential.



Compared to what, Joe?

- How well you think you can do?
- How well you think you should do?
- How other students in the class did?
- Your parent's opinion of what you should do?
- Mr. Cgull's opinion of what you can do?
- Your friends' opinions of what you should do?
- The requirements to get into Harvard?
- The requirements to get into Community college?
- The requirements to get a job at Kodak?

Guidance and Counseling. Each student deserves to know about his or her performance on a test.

The meaning of test results to the student (and parents and teachers) depends upon the grade level, the specific test, and the purposes for taking the test. In lower grade levels, counseling would primarily involve skill development during study in interest areas while expanding one's experience background. In higher grade levels counseling becomes more specific in relation to course of study and career decisions. Information from standardized tests, through judicial interpretation, can help the student better understand his relationship to group norms. Remember, the norms are not goals to be attained!

USES OF GROUP DATA

The most reliable information from standardized testing consists of statements about the group. Individual variations, those who score high and those who score low, tend to balance out providing relatively stable insights into the group as a whole.

Group statistics of central tendency, means, medians, **standard deviations,** quartiles, percentiles, etc., are particularly useful for program analysis and evaluation. Comparing the results achieved by one group with those of another is the most common use of group data. Ms. Jones checks to see if her class achieved as well as Ms. Smith's class or as well as students collectively across the school district or the state. Mr. Brown compares his students' achievement against "national" figures for the group upon which the test was normed. The **validity** of such comparison depends upon the similarity of characteristics of the groups being compared and on the nature of the interpretative conclusions drawn.

These are not the only uses of group data - - item analysis, comparisons of group achievement across subtest content areas, and correlation with other test results on the same content provide useful information.

Group statistics. Interpreting the meaning of descriptive statistics requires knowing one's expectations as well as one's status. That the median score for class A is higher than the median for class B only tells relative status. Is that good? Bad? Irrelevant? This depends on the two groups. If class A is a homogeneous group of "low achievers" and class B is a so called "normal class," then such a comparison would be unexpected and the stimulation for investigation of cause. If the results were reversed, it would be compelling to say "that's to be expected!" But, how much lower should the average of a "low" class be when compared to a "normal" class? How much difference would be significant? It is important to realize that significance is someone's value judgment. Even when choosing statistical significance levels, someone must decide the degree to which difference will be considered significant. And, if two scores differ significantly on a statistical basis, this still may not be practically significant. That is, while real, the differences may not be of a magnitude to warrant changes in programs, expenditures, behaviors, etc.

The median, quartiles and percentiles. When one group is compared to another by the use of medians, we know only whether the middle score (median) of group A is higher or lower than that of group B. Quartiles and percentiles add more information about the group. A brief study of figure 5.8 shows that the two groups have the same median score, yet the other data available show the groups to be quite different. Ninety percent of class A attained scores of 96 or lower, that is, 10% scored above 96, whereas

STANDARD SCORES - READING

	Class A	Class B
Highest score	100	122
90%ile	96	109
Q3 or 75%ile	92	101
Q2 or median or 50%ile	83	83
Q1 or 25%ile	76	65
10%ile	68	54
Lowest score	66	38
Standard deviation	6.7	12.8

By comparing descriptive statistics, one can learn how groups are similar and different.

Figure 5.8

in class B, 10% scored higher than a standard score of 109. Class B is thus seen to have more higher achieving students than class A. Likewise class A has fewer low achieving students. Class A is more homogeneously clustered (the middle 80% of the class fell within a range of 28 points) than class B (where the middle 80% of the class spread over a range of 55 standard score points). Twenty-five percent of class B had scores of 65 or lower, whereas the lowest score in class A was 66. Obviously, instruction for the two groups must be different if the individuals in the groups are to have their needs met.

The standard deviation is a concise description of how spread out or clustered are the scores of a group. The larger standard deviation of class B tells that these scores spread over a larger range than do the scores of class A.

Item analysis.

Going beyond examination of group summary statistics yields valuable information. Item analysis is a technique for a program evaluation which looks at results on a group basis, but very specifically. For each test item, a tally is made to determine how many (what percent of) students correctly responded to the item. Knowing how well students performed on particular sets of related questions allows the teacher to identify areas of strength and/or weakness.

Item analysis information may be obtained on teacher made tests by hand tally or by machine scoring if student responses are made on machine score answer sheets. On standardized tests that are scored by machine, item analysis information may be obtained as part of the processing. Item analysis formats vary depending upon what scoring service prepares the results, but regardless of format, the same information is obtained and is used primarily for program examination.

Two examples of item analysis formats are shown below in figures 5.9 and 5.12.

EXAMPLE 1 - ITEM ANALYSIS FORMAT

Reading — Grade 5

Question No.	Number of Students Responding: No Response	Response Number 1	2	3	4	5	Percentage Response: No Response	Response Number 1	2	3	4	5
1			1		23*		.0	.0	4.2	.0	95.8*	.0
2			1	23*			.0	.0	4.2	95.8*	.0	.0
3		5	19*				.0	20.8	79.2*	.0	.0	.0
4		12	8*	2	2		.0	50.0	33.3*	8.3	8.3	.0

Figure 5.9

In order to use such information, it is necessary to know what each question is about. Through use of the MAT Content Outline (or examiniation of the test if a breakdown is not available), the questions pertaining to certain topics or skill areas can be identified within each subtest. A question or group of questions can be looked at to see how a class or an entire grade level responded.

What information does this give? In example 1 (Fig. 5.9) above, 95.8% scored correctly on item one; on item three, 79.2% scored correctly with all of those responding incorrectly choosing distractor (1); on item four, more students chose a single wrong answer (50% chose A) than answered correctly (33.3%). How does one use this information? Item one - no problem! Item three - not bad! Item four - concern! Students must have a misconception that can be corrected!!! Or, was this just a particulary difficult question? How well did students in the normative population achieve on this item? From a list of item

Test 2: READING (45 items)

Topic	Item Numbers Form F	Form G	Form H
Main Thought	5, 11, 18, 24, 31, 37, 43	7, 11, 19, 25, 29, 33, 41	1, 8, 15, 17, 23, 29, 34, 41
Inferential	3, 4, 8, 10, 14, 16, 20, 21, 22, 28, 30, 33, 35, 36, 40, 41, 44, 45	1, 3, 8, 9, 10, 12, 13, 16, 21, 24, 27, 31, 32, 34, 36, 39, 40, 42	3, 4, 7, 10, 12, 14, 19, 22, 24, 26, 27, 30, 32, 33, 36, 38, 39, 43, 44, 45

Figure 5.10

METROPOLITAN ACHIEVEMENT TESTS
ITEM DIFFICULTY VALUES

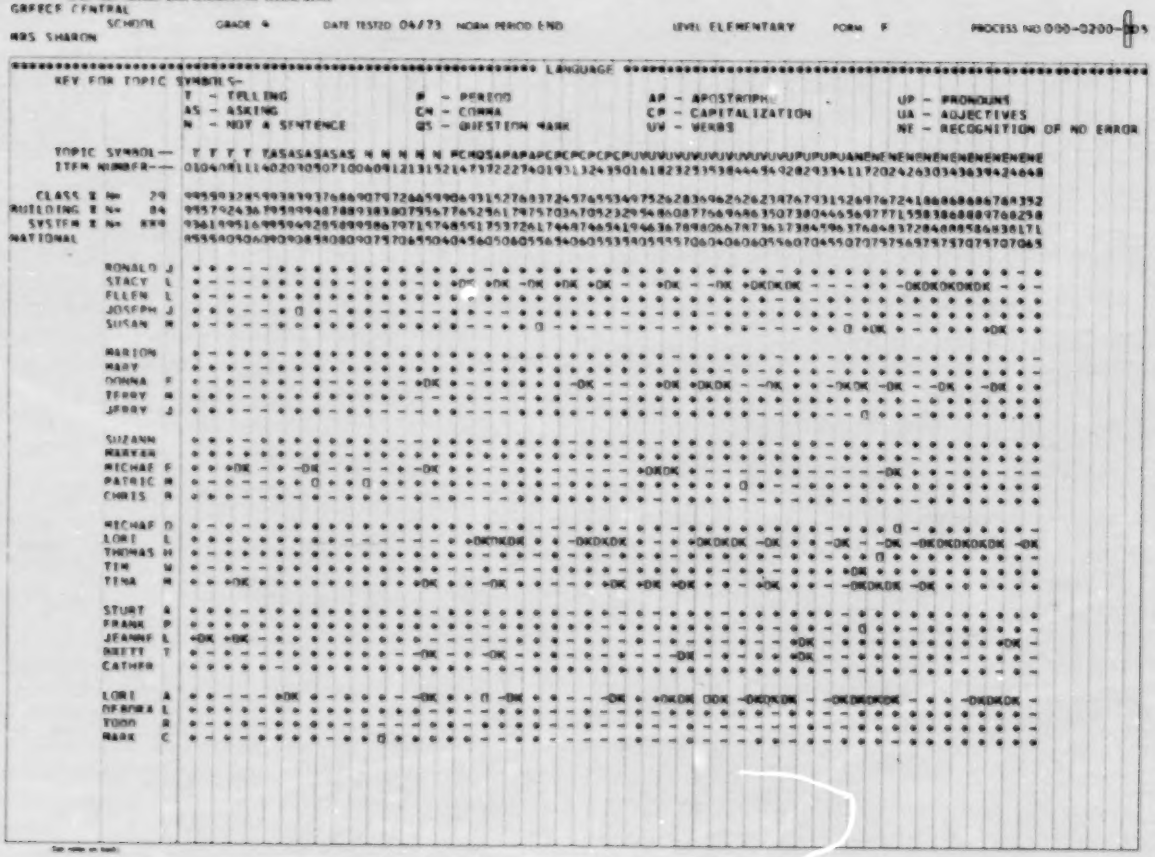
Intermediate			Form F			End of Grade 5	
Item No.	Word Know.	Read-ing	Lan-guage	Spel-ling	Math Comp.	Math Conc.	Math Prob. Solv.
1	90	85	80	95	90	80	90
2	85	85	70	90	90	85	80
3	95	55	50	85	85	80	75
4	90	40	60	85	80	90	70
5	85	65	60	70	85	65	75
6	90	80	55	90	80	70	75
7	75	85	55	80	75	65	65

Figure 5.11

difficulty values (Fig. 5.11) provided by the publisher, we find that 40% of the students in the norm group scored correctly on this item. Our concern decreases as we place our low achievement on this item into perspective. But, how do our students do on other questions related to this skill? Consulting the content outline (Fig. 5.10) we find several questions that along with item 4, measure inferential skills. Do results on these questions show weakness in this general area? Suppose this is the case, what should be done? Since inferential skills are important in math, science and social studies, as well as in reading comprehension, several opportunities exist for "working on" this skill area. Being aware that attention is needed to this skill area is the initial phase of program adjustment. Standardized test data have done their job! No great catastrophic impact, just another insight through the use of this particular tool!

On example No. 2, each student's response to each question of the subtest is listed as being correct (+), incorrect (-), no response (0), or in some areas such as math, a don't know (DK) response may be indicated. With this particular type of format, you cannot tell how students responded to the various distractors for each item. Otherwise the use of information is substantially the same as with the example 1 format.

EXAMPLE 2 - ITEM ANALYSIS FORMAT



METROPOLITAN ACHIEVEMENT TESTS — ITEM REPORT

GREECE CENTRAL — SCHOOL — GRADE 4 — DATE TESTED 04/73 — NORM PERIOD END — LEVEL ELEMENTARY — FORM F — PROCESS NO 000-0200-[illegible]

MRS SHARON

LANGUAGE

KEY FOR TOPIC SYMBOLS—

T – TELLING; P – PERIOD; AP – APOSTROPHE; UP – PRONOUNS
AS – ASKING; CM – COMMA; CP – CAPITALIZATION; UA – ADJECTIVES
N – NOT A SENTENCE; QS – QUESTION MARK; UV – VERBS; NE – RECOGNITION OF NO ERROR

TOPIC SYMBOL— T T T T TASASASASAS N N N N N PCMQSAPAPAPCPCPCPCPCPUVUVUVUVUVUVUVUVUPUPUPUANENENENENENENENENE

ITEM NUMBER— 0104081111402030907100609121331521473722227401931324350161823253538444549282933411720242630343639424648

CLASS % N= 29 — 9959937855993839376869079726855990493152763372697655349752628359626262387670315269762741868868867683 92

BUILDING % N= 84 — 9957924367959994878893838079956776529617975703470523295948087766946635073804465697771558386880976825 8

SYSTEM % N= 899 — 9361995169959692958999867971547855175372617448746541946367898067877361738459637468372840885866381 7L

NATIONAL — 9555909069909085908090757065904045605060556940605539405555706040606055960704550707575697975707797070065

RONALD	J	+ + + – – + + + + + + + + + + + + + + + + – + + + + + + + + + + + + + + + +

At the top of example 2, (Fig. 5.12) reference data are shown; these include national percentages of correct responses, district percentages and the class percentages. Items are grouped by skill areas as is done in the content outline; this is shown by use of topic symbols and column separation between the grouped items. By using the topic symbols to locate the questions in the subtest you want to look at, you can see how a class performed on those items and compare the results against national and district results.

A class with 35% responding correctly to a question with a difficutly value (percentage achieved in the national normative group) of 30% would not be a major concern. If, however, 30% of a class responded correctly to an item with a difficulty level of 80%, you would be led to look for reason why this occured.

Group achievement across subtests.

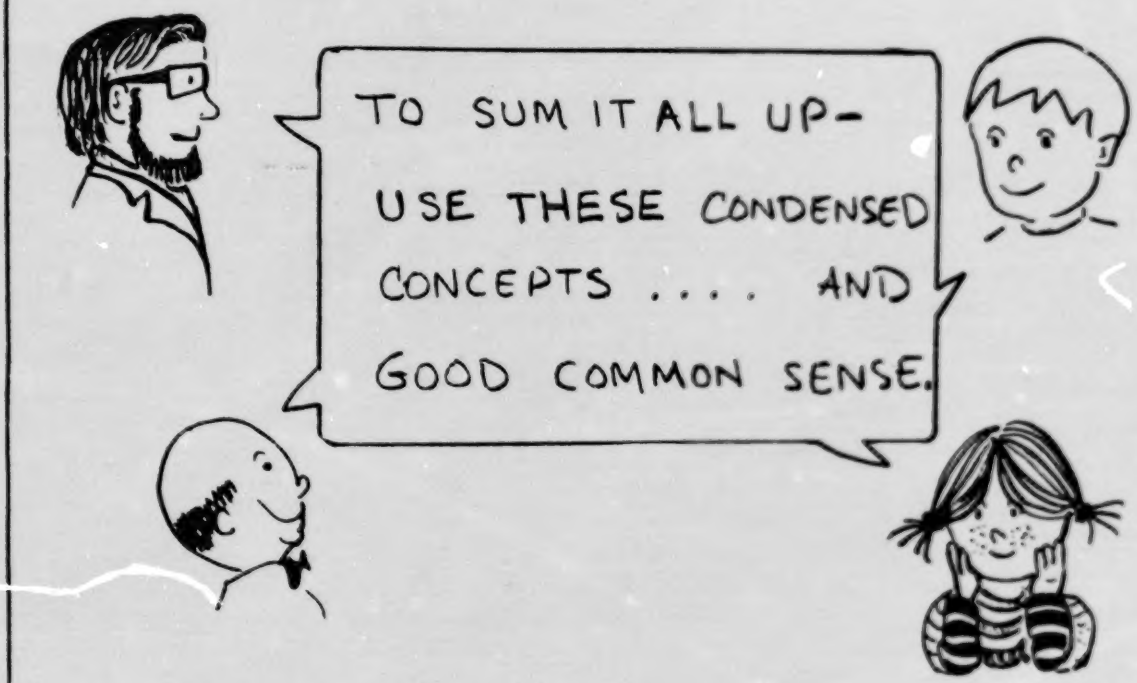
In a similar, although less detailed, approach to item analysis, subtest content areas may be compared to identify strengths and weakness. This procedure on a group basis is very similar to that described under Uses of Individual Data, except that here the impact is upon program rather than any single student. If, for example, a groups' scores show consistent approximation of normative scores in all areas except spelling, which is much lower, it is probable that some feature of the local spelling program, emphasis, or method is related to the low standing. Such an identification, and raising the question "why?", is the work of standardized testing.

Correlation with other data.

Before the results from a standardized test become the impetus for drastic change, they should be related to other information. This might be by fairly sophisticated statistical techniques, or simply by subjective synthesis of other indicators. Nothing as complex as a curriculum is dependent upon a single causative feature, and should not be expected to be described by a single score or set of scores from a standardized test. A test score, in concert with other scores, grades, teacher comments and judgments, student attitude, etc. can and should provide a sound basis for evaluating the effectiveness of programs.

■■■

SUMMARY


TO SUM IT ALL UP-
USE THESE CONDENSED CONCEPTS .... AND GOOD COMMON SENSE.

SUMMARY

This section consists of brief statements of ideas chosen because of their cogency and power. These concepts are discussed in some detail in the text of this handbook and in a multitude of books, articles, and papers relating to this topic.

Numeration of the following points relates them contextually to handbook sections.

1.1 Testing is a technique for obtaining information; it is a tool.

1.2 Data from standardized testing can be effectively used and woefully misused.

1.3 People (students), while conceptually equal under the law, differ in **ability**, interests, competencies, creativity, sensitivity, etc. Learning theory must be based upon the inherent differences among people.

1.4 Human diversity should be held in high regard; it's all right to be different.

2.1 District goals and philosophy are an amalgam from all who comprise the district: parents, teachers, students, administrators, support staff, taxpayers.

2.2 Evaluation in the broad sense is a component of decision making.

2.3 Evaluation is a process which affects the situations being evaluated.

2.4 Standardized testing is only one means of evaluation; there are many important educational objectives that are not measured by **standardized tests.**

2.5 Evaluative feedback is important for evolving programs.

2.6 Although expected to do many other things, schools are primarily charged with the responsibility of bringing students to proficiency in the use of basic skills of communication--"readin', writin' 'n' rithmetic."

2.7 All behaviors should manifest a regard for the dignity and value of each individual.

2.8 Data, relevant and valid, are important to decision making.

2.9 Standardized tests (norm referenced) differ in purpose from assessment tests (criterion referenced).

2.10 Standardized **achievement test** data should affect, but not determine, the curriculum.

2.11 Normative, or average, performance descriptions are reference points, not goals.

3.1 Test data should be gathered to serve a clearly defined purpose.

3.2 The "standards" of standardized tests are simply regulations for controlling variables other than student competency.

3.3 Standardized test data are not appropriately precise for use in showing short term learning gains.

4.1 Each person to be tested is entitled to an environment and preparation that will optimize the opportunity for success.

4.2 Both the person being tested and the person administering the test should understand the purpose for testing and how results will be used.

4.3 Directions for test administration should be followed exactly to assure uniformity and thus comparability of results.

4.4 After testing is completed, materials should be organized, sent for scoring, and preparation made for return of results.

5.1 The basis upon which standardized tests are constructed is the comparison of results between populations. They are designed to differentiate students one from another in terms of achievement on the test.

5.2 **Norms** are statements of the average **achievement** of some referent population.

5.3 A test score merges all of the various responses to questions into a single value.

5.4 Nothing as complicated as the mental processes of a human being can be expressed as a test score.

5.5 Test scores and related descriptive statistics have various characteristics, strengths and limitations, that must be understood in order to use such values appropriately.

5.6 **Ability test** scores should not be considered as the comparative standard against which to relate other scores, but simply as another measurement value to be weighed along with others.

5.7 In analyzing test results, one should look not only for the unusual, but also for the patterns that are most representative - - the usual.

5.8 The most valid use of group administered tests is for making statements about the group (as opposed to individuals).

5.9 The primary use of standardized test results is for program evaluation and evolution.

Pervading all of the behaviors and ideas related to instruction and the use of testing information must be a regard for the affective. Whatever is done in the interactions among people involves feelings and affective impact. People are tender and should be "handled with care." This is especially true of students who are in such formative phases of their lives.

Students' test scores should be treated with respect for the rights of each individual. This includes a person's right to know information about himself as well as feeling assured that information is treated confidentially.

Test results should be viewed with balanced concern. It is easy to be indifferent to information derived from testing by enumerating its limitations. It is also easy to overemphasize small differences because they show up in statistical data.

In our culture it is common to seek causes for observed effects. It is, however, very difficult to validly assign causation, particularly in situations where many factors contribute to the effect. There is a great temptation to assume causality simply because two things are related, especially where the two correlate highly. Failure to recognize that correlation does not mean cause and effect leads to many unwarranted conclusions.

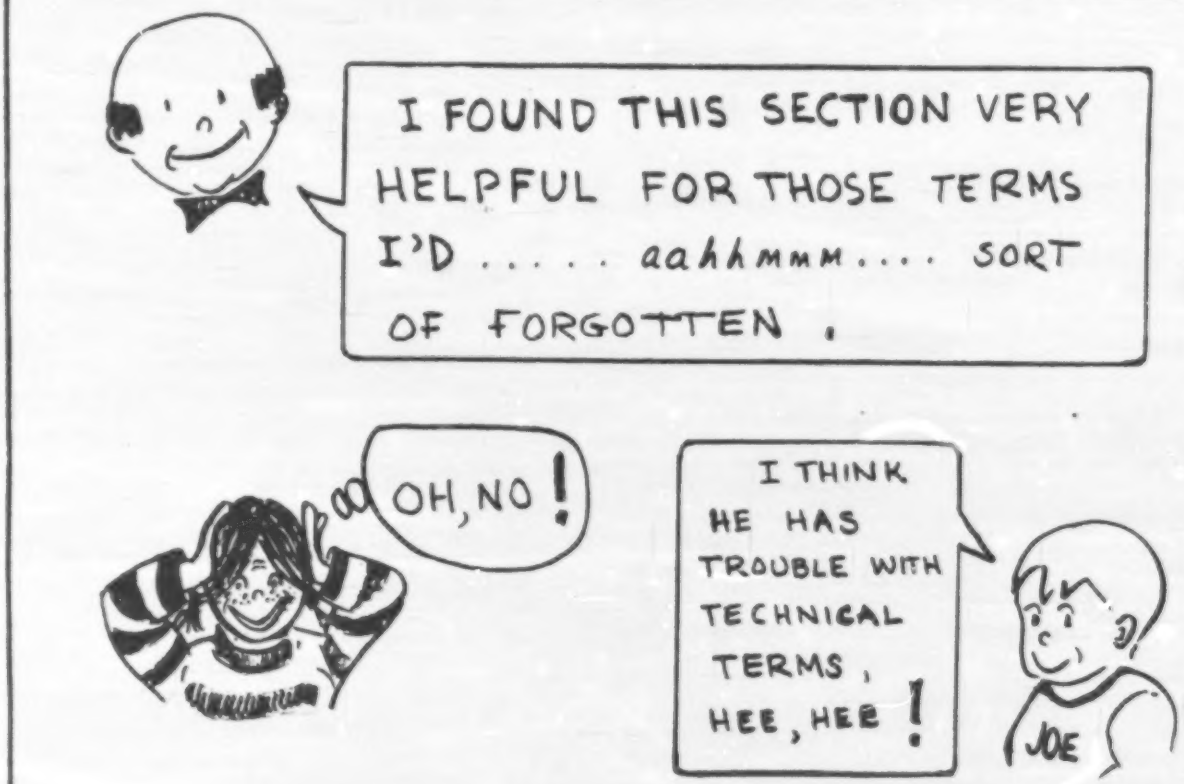
"Objective" data somehow are often regarded as more valid, more potent, than subjective information simply because they are quantified. Unquestioning reverence for objective data is not warranted. Objective data are useful,

easy to manipulate and relate, but they are not inherently better than subjective experienced wisdom or insight.

Test results answer few questions. New questions at more precise levels are raised. Thus interpretation of test results and translation into action must be done judiciously and deliberately. Seven hundred years ago, Roger Bacon, the English scientist and philosopher, expressed what is applicable today "Crafty men condemn studies, foolish men admire them; wise men use them." Testing, as a means of study, is a tool to be used.

■■■

GLOSSARY


I FOUND THIS SECTION VERY HELPFUL FOR THOSE TERMS I'D ..... aahhmmm.... SORT OF FORGOTTEN .
OH, NO !
I THINK HE HAS TROUBLE WITH TECHNICAL TERMS, HEE, HEE !
JOE

GLOSSARY OF TERMS

ACHIEVEMENT: An indication of accomplishment of a skill or an activity.

ACHIEVEMENT TEST: A test that measures the extent to which a person has "achieved" something, acquired certain information, or mastered certain skills - usually as a result of planned instruction or training.

AGE NORMS: Scores representing the typical or average performance for persons in successive age groups. Such norms are generally used in the interpretation of mental ability test scores.

ANALYSIS: The process of separating or breaking up any whole into component parts so as to find out their nature, proportion, function, relationship, etc. Item analysis breaks down a test by questions, class analysis breaks down the class by grouping students' scores. Analysis is the opposite of synthesis, wherein individual components are observed and combined to form an overall view not apparent or complete simply be seeing each part.

APTITUDE: A combination of abilities and other characteristics, whether native or acquired, that are an indication of an individual's capacity to learn or to develop proficiency in some skill or subject matter if appropriate training is provided.

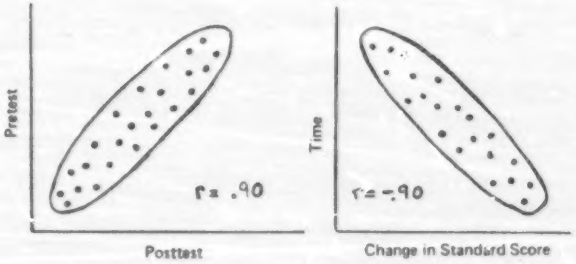
APTITUDE TEST: This type of test includes those of general academic ability, and those of special abilities such as verbal, numerical, mechanical, musical, etc. which measure both ability and previous learning and are used to predict future performance.

AVERAGE: A general term applied to the various measures of central tendency. The three most widely used averages are the mean, the median, and the mode. When the term "average" is used without designation as to type, the most likely assumption is that it is the mean.

BIVARIATE CHART: A diagram in which a tally mark is made to show the scores of one individual on two variables. The intersection of lines determined by the horizontal and vertical scales form cells in which the tallies are placed. Such a plot provides frequencies for the two distributions, and portrays the relation between the two variables. This gives a visual picture of the correlation between the two variables and serves as the basis for computation of a correlation coefficient. (See Figure 5.7).

CLASS LIST: An alphabetical list report of pupils' scores on the various subtests taken. Pupils' scores can be grouped by class, or by the total grade, whichever is preferred. (See figures 5.1 and 5.2.)

CORRELATION COEFFICIENT: An expression of the degree of relationship between two sets of measures for the same group of individuals. Correlation coefficients range from .00, denoting a complete absence of relationship, to +1.00 and to -1.00, indicating a perfect positive or negative correspondence, respectively. The letter r is commonly used to refer to the correlation coefficient.



CRITERION-REFERENCED TEST: A test designed to provide information on the specific knowledge or skills possessed by a student. Such tests usually cover relatively small units of content and are closely related to instruction. Their scores have meaning in terms of what the student knows or can do, rather than in their relation to the scores made by some external reference group.

DEVIATION IQ (DIQ): An age-based index of general mental ability. It is based on the deviation or difference between a person's obtained score and the score that is average for persons of that chronological age. DIQ is the commonly used "IQ" score, although it is technically different in derivation from an actual IQ (non-deviation value).

DIAGNOSTIC TEST: A test used to locate an individual's specific areas of weakness or strength, to determine the nature of his weakness or deficiencies, and to suggest their cause whenever possible.

DIFFICULTY VALUE: An index which indicates the percent of some specified group, such as students of a given age or grade, who answered a test item correctly. It can be used with an item analysis report as reference data.

ERROR OF MEASUREMENT: As applied to a single obtained score, the amount by which the score may differ from the hypothetical true score due to errors of measurement. The larger the standard error of measurement, the less reliable the score. The standard error of measurement is an amount such that in about two-thirds of the cases, the obtained score would not differ by more than one standard error of measurement from the true score.

FORMATIVE: Evaluations used to determine the degree of mastery of a given goal or learning task and to pinpoint the part of the goal or task not attained or mastered. Evaluations and tests used for formative purposes are part of a process; they help to form subsequent steps in the process. This differs from summative usages (which sum up) that come at the end of a course and serve mostly for grading or goals-met types of evaluation.

FREQUENCY DISTRIBUTION: A tabulation of scores from high to low showing the number of individuals that obtained each score or fell in each score interval. (See figures 5.3 and 5.4.)

GRADE EQUIVALENT (GE): The grade interpretation of the relationship of a given raw score to the standardization population of a particular test expressed in terms of the grade and the month within that grade level. The most valid use of GE's is in the interpretation of the median performance of a group of students.

GRADE NORMS: Scores representing typical or average performance for persons of various grade placements.

INTELLIGENCE QUOTIENT (IQ): Originally, an index of brightness expressed as the ratio of a person's mental age to his chronological age (MA/CA), multiplied by 100 to eliminate the decimal. This quotient IQ has been gradually replaced by the deviation IQ concept (DIQ).

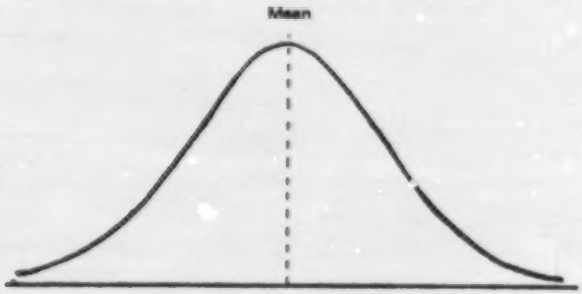
MASTERY TEST: A test designed to determine whether a pupil has mastered a given unit of instruction or skill.

MEDIAN (MD): The middle score in a distribution or set of ranked scores; the 50th percentile.

MODE: The score (or value) that occurs most frequently in a distribution. It is possible to have more than one mode in a distribution of scores.

N: The symbol commonly used to represent the number of cases in a group.

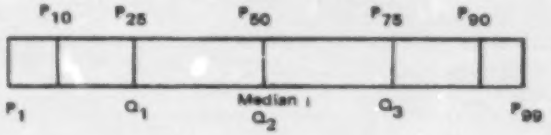
NORMAL DISTRIBUTION: A distribution of scores or measures that in graphic form has a distinctive bell-shaped appearance. In such a normal distribution, scores or measures are distributed symmetrically about the mean. Cases are concentrated near the mean and decrease in frequency the farther one departs from the mean.



Normal Distribution

NORMS: Statistics that supply a frame of reference by which meaning may be given to obtained test scores. Norms are based upon the actual performance of pupils of various grades or ages in the standardization group for the test. Since they represent average or typical performance, they should not be regarded as standards or universally desirable levels of attainment.

PERCENTILE (P): A point (score) in a distribution at or below which fall the percent of cases indicated by the percentile. A score coinciding with the 60th percentile (P60) is regarded as equaling or surpassing that of 60 percent of the persons in the group, with 40 percent of the performances exceeding this score. A student's converted score indicating his relative position in a specific group may be expressed as his percentile rank.



PERMANENT RECORD LABEL (PRESSCORE LABEL): This gummed-back or stick-on label shows the pupil's name and his scores for each test (subtest) taken. This type of report is especially helpful for administrative records since it precludes the need for hand-copying of pupil scores into cumulative records.

QUARTILE (Q): One of three points that divide the cases in a distribution into four equal groups. The lower quartile (Q_1), or 25th percentile, sets off the lowest fourth of the group; the middle quartile (Q_2), the 50th percentile which is the median, divides the second fourth of cases from the third; and the third quartile (Q_3), 75th percentile, sets off the top fourth.

RANDOM SAMPLE: A sample of the members of some total population drawn in such a way that every member of the population has an equal chance of being included. It is done in such a way that precludes the operation of bias or selection.

RANGE: For some specified group, the difference between the highest and the lowest score on a test.

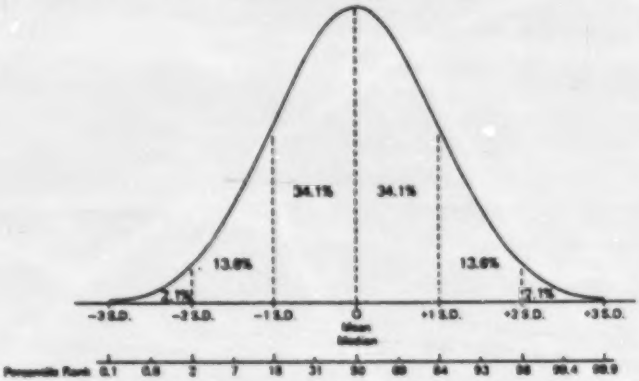
RANK LIST: A list report of pupils, ranking the pupils from the highest score to the lowest score, based on a specific criterion score, e.g., reading, math, etc. This could be done using raw scores, standard scores, stanines, etc.

RAW SCORE: The first quantitive result obtained in scoring a test. Usually the number of questions answered correctly or the number right minus some fraction of the number of wrong answers.

RELIABILITY: The extent to which a test is consistent in measuring whatever it does measure; dependability, stability, trustworthiness, relative freedom from errors of measurement.

REPRESENTATIVE SAMPLE: A sample that corresponds to or matches the population of which it is a sample with respect to characteristics important for the purposes under investigation. In an achievement test norm sample, such significant aspects might be the proportion of cases of each sex, from various types of schools, different geographical areas, several socioeconomic levels.

STANDARD DEVIATION (S.D.): A calculated value which expresses the way a group of scores spread out, or are clustered around the mean. In a normal distribution, plus and minus one standard deviation from the mean sets off approximately 2/3 of the population. The smaller the S.D., the more homogeneous the group being examined.

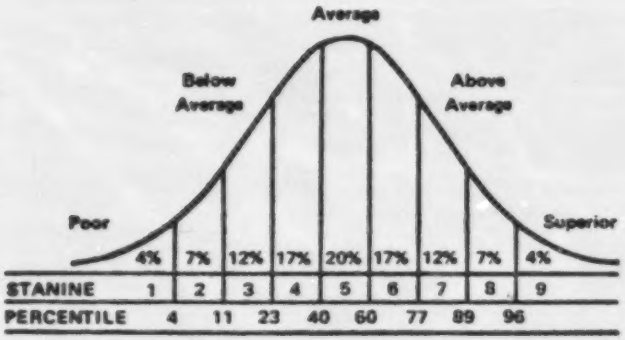


Standard deviation units measure off areas of the normal curve. Relationship of percentiles is also shown.

STANDARD SCORE: A general term referring to a score distribution which has been "standardized" to a specific mean and a specific standard deviation. This allows raw scores to be converted to standard scores for convenience, comparability, and ease of interpretation.

STANDARDIZED TEST: A test designed to provide a systematic sample of individual performance, administered according to prescribed directions, scored in conformance with definite rules, and interpreted in reference to certain normative information.

STANINE: One of the steps in a nine-point scale of standard scores. The stanine (short for standard-nine) scale ranges from 1 to 9 with a mean of 5 and a standard deviation of 2.



Distribution of stanines in a normal population showing relationship to percentiles.

SUMMATIVE: Evaluations used as a general assessment of the degree to which goals or learning tasks have been attained over an entire program or course, or some substantial part of it. (See FORMATIVE.)

VALIDITY: The extent to which a test does the job for which it is used.

■■■

NO WONDER MR. CGULL HAS TROUBLE WITH SOME OF THOSE TERMS!
YEAH, AND REMEMBER WHEN YOU THOUGHT OUR TEST QUESTIONS WERE SO HARD....!